1. Diagnostic Errors

Introduction

Background

Diagnostic error, as defined by the National Academy of Medicine in 2015, is "the failure to (a) establish an accurate and timely explanation of the patient's health problem(s) or (b) communicate that explanation to the patient."[1] This definition focuses on the outcomes of the diagnostic process, recognizing that diagnosis is an iterative process that solidifies as more information becomes available. The diagnosis needs to be timely and accurate so that appropriate treatment is initiated to optimize the patient's outcome. Any gaps that arise in the diagnostic process can lead to error. In this chapter we discuss four patient safety practices (PSPs) that have the potential to decrease diagnostic errors: the use of clinical decision support (CDS); result notification systems (RNS); education and training; and peer review.

Importance of Harm Area

Diagnostic error is an increasingly recognized threat to public health, with estimates of 5 percent of adults being affected in the outpatient environment.[2] In the hospital setting, diagnostic error is responsible for 6 to 17 percent of adverse events.[1,3] Diagnostic error has also been shown to be responsible for more closed malpractice claims than other causes.[1,4,5] The Institute of Medicine (now the National Academy of Sciences), in their seminal report on diagnostic safety, concluded that "most people will experience at least one diagnostic error in their lifetime."[1]

PSP Selection

Using systematic reviews and reports, the Technical Expert Panel, Advisory Group, and Agency for Healthcare Research and Quality developed and reviewed an initial list of 23 PSPs that target diagnostic errors. Studies have uncovered two broad categories of underlying root causes: cognitive-based factors, such as failed heuristics, and systems-based factors, such as lack of provider-to-provider communication and coordination.[2,6,7] Therefore, the PSPs selected by consensus for inclusion in this report addressed one or both of these fundamental high-leverage areas.

- **CDS** offers solutions integrated into the workflow to address diagnostic errors by providing stakeholders with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to improve decision making and communication.[8]
- **RNSs** aim to address lapses in communication, a contributing factor to delayed diagnosis and treatment of patients in both ambulatory and inpatient settings.[9,10]
- **Education and training** on the diagnostic process enhance clinical reasoning and decrease biases.[6]
- **Peer review** identifies potential diagnostic errors before they reach the patient and provides feedback with the intent of improving clinical practice and quality.[1,11]

References for Introduction

1. National Academies of Sciences, Engineering and Medicine. 2015. Improving Diagnosis in Health Care. Washington, DC: National Academies Press. https://doi.org/10.17226/21794.
2. Singh H, Meyer AN, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf. 2014;23(9):727-31.
3. Zwaan L, de Bruijne M, Wagner C, et al. Patient record review of the incidence, consequences, and causes of diagnostic adverse events. Arch Intern Med. 2010;170(12):1015-21.
4. Tehrani ASS, Lee H, Mathews SC, et al. 25-Year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the National Practitioner Data Bank. BMJ Qual Saf. 2013;22(8):672-80.
5. Schiff GD, Puopolo AL, Huben-Kearney A, et al. Primary care closed claims experience of Massachusetts malpractice insurers. J AM Med Assoc Intern Med. 2013;173(22):2063-8.
6. Graber ML, Kissam S, Payne VL, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ Qual Saf. 2012;21(7):535-57. doi:10.1136/bmjqs-2011-000149.
7. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care—a systematic review. Fam Pract. 2008;25(6):400-13.
8. Official Website of The Office of the National Coordinator for Health Information Technology (ONC). https://www.healthit.gov/.
9. Callen JL, Westbrook JI, Georgiou A, et al. Failure to follow-up test results for ambulatory patients: a systematic review. J Gen Intern Med. 2012;27(10):1334-48.
10. Davis Giardina T, King BJ, et al. Root cause analysis reports help identify common factors in delayed diagnosis and treatment of outpatients. Health Aff. 2013;32(8):1368-75.
11. Butler GJ, Forghani R. The next level of radiology peer review: enterprise-wide education and improvement. J Am Coll Radiol. 2013;10(5):349-53.

1.1 Patient Safety Practice: Clinical Decision Support

Authors: Kendall K. Hall, M.D., M.S., Eleanor Fitall, M.P.H., and Kristen Miller, Dr.P.H.

Reviewer: Katharine Witgert, M.P.H.

1.1.1 Practice Description

Key Findings:

- CDS has been shown to improve diagnosis in exploratory and validation studies, but the tools need to be fully implemented and tested in clinical settings.
- CDS is best used as an adjunct to the clinician's decision-making process and not as a replacement.
- The diagnoses generated by CDS tools are only as good as the information that is put into the system; if the initial assessment of the patient (e.g., physical exam finding) is incorrect, the output is likely to be incorrect.
- Despite their potential, diagnosis generators have had limited use, owing in large part to challenges integrating them into busy clinicians' workflows.

Diagnostic error is a complex and multifaceted problem that requires systems solutions to achieve the necessary changes. Advancements in health information technology (IT) represent thoughtful and sophisticated ways to reduce delayed, missed, or incorrect diagnoses.[1] Contributions of health IT include more meaningful incorporation of evidence-based diagnostic protocols with clinical workflow, and better usability and interfaces in the electronic health record (EHR).

The Office of the National Coordinator for Health Information Technology defines CDS as providing "clinicians, staff, patients or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and healthcare. CDS encompasses a variety of tools to enhance decision making in the clinical workflow. These tools include computerized alerts and reminders to care providers and patients; clinical guidelines; condition-specific order sets; focused patient data reports and summaries; documentation templates; diagnostic support, and contextually relevant reference information, among other tools."[2]

CDS represents a range of different interventions, from documentation templates to interruptive popup alerts. The knowledge bases triggering CDS differ as well. Rules-based or logic-based CDS often takes the form of IF-THEN rules. More advanced CDS leveraging artificial intelligence (AI) and machine learning taps awareness of past experiences and patterns in clinical data. These techniques have generated interest and excitement in their potential to better augment clinician intelligence and support decision making.

Several patient safety researchers have suggested that health IT, including CDS, can be leveraged to improve diagnosis, although the data have been mixed.[1,3-7] Therefore, the question of interest for this review is, "Does CDS lead to improved diagnostic performance?" This review's key findings are located in the box above.

1.1.2 Methods

We searched four databases (CINAHL®, MEDLINE®, PsycINFO®, and Cochrane) for articles published from 2008 through 2018 using the terms "diagnostic errors," "delayed diagnosis," "missed diagnosis," and their synonyms. Terms specific to this PSP include "clinical decision support," "medical informatics applications," "artificial intelligence," "computer-aided decision making," "computer-assisted diagnosis," and related terms. The initial search yielded 2,208 results. Once duplicates had been removed and additional relevant referenced articles added, a total of 2,202 articles were screened for inclusion, and 87 full-text articles were retrieved. Of those, 37 studies were selected for inclusion in this review.

Articles were excluded if they were not focused on use of CDS specifically for diagnosis (e.g., focus on use of CDS for medication ordering), the outcome was not relevant to this review, the article was out of scope, or the study was of significantly limited rigor.

General methods for this report are described in the Methods section of the full report.

For this patient safety practice, a PRISMA flow diagram and evidence table, along with literature-search strategy and search-term details, are included in the report appendixes A through C.

1.1.3 Evidence Summary

1.1.3.1 CDS To Generate Diagnoses

1.1.3.1.1 Differential Diagnosis Generators

Differential diagnoses (DDX) are a list of diagnostic hypotheses generated by the clinician during the course of the patient interaction, and are based on information such as the history and physical exam. Often several different diagnostic possibilities are initially present, and as the clinician gathers additional information to support or refute the hypotheses, the list can be narrowed until arriving at the correct diagnosis.

DDX generators are "programs which assist healthcare professionals in clinical decision making by generating a DDX based on a minimum of two items of patient data."[8] DDX generators provide a list of potential diagnoses for consideration, sometimes in order of likelihood based on available information, as a means to improve diagnosis.

The first study discussed is a systematic review and meta-analysis conducted by Riches et al. (2016), which included 36 articles investigating the effects of 11 different DDX generators to retrieve accurate diagnoses (i.e., the correct diagnosis appeared in the list of possible diagnoses). Of note, only five of the tools are still in existence. Using different computational approaches, such as pattern matching and Bayesian probabilities, these diagnostic aids generate lists of DDX for consideration based on clinical data that the user inputs. With respect to the effectiveness of the DDX generators at retrieving accurate diagnoses, the authors concluded that the pooled accurate diagnosis retrieval rate was high, although with considerable heterogeneity (pooled rate=0.70, 95% confidence interval [CI], 0.63 to 0.77; I^2 = 97%, p<0.0001). In the subgroup analyses examining the accuracy of individual DDX generators, ISABEL, one of the tools under evaluation, outperformed all of the other tools, but again, the heterogeneity was considerable (pooled rate = 0.89, 95% CI, 0.83 to 0.94; I^2 = 82%, p<0.0001). When comparing the performance of the DDX tools to that of clinicians, the authors found that the DDX tools were associated with a small, nonsignificant increase in accurate diagnosis retrieval.[8]

In a study by David et al. (2011), the primary objective was to determine the misdiagnosis rate of cellulitis, an infection of the skin and tissue underneath, but the authors also determined whether or not a visually based, computerized diagnostic decision support system (VCDDSS) could generate an improved DDX based on the presenting signs and symptoms for the misdiagnosed patients. The system requires the user to input relevant patient findings (e.g., clinical information, physical examination findings) to generate a ranked list of potential diagnoses. Using a cellulitis-specific module of the VCDDSS, the authors found that the system included the correct diagnosis in the DDX 64 percent of the time. This was significantly greater than the diagnostic accuracy of the admitting residents, who included the correct diagnosis in their DDX only 14 percent of the time without the use of the VCDDSS (p=0.0003).[9] Gegundez-Fernandez et al. (2017) evaluated the diagnostic performance of Uvemaster, a mobile DDX generator that provides a ranked list of syndromes that cause uveitis, a form of eye inflammation, based on clinical findings. The percentage of cases for which a diagnosis included in the DDX by the app matched the original clinician diagnosis was 96.6 percent (95% CI, 84.1 to 96.6). When the diagnoses were ordered by sensitivity, the original diagnosis was listed within the top three diagnoses generated by the app in 90.9% of cases (95% CI, 84.1 to 96.6) and was listed as the first diagnosis in 73.9% of cases (95% CI, 63.6 to 83.0).[10]

Using real-case vignettes, Segal et al. (2014) and Segal et al. (2016) both evaluated the use of a DDX generator, SimulConsult, on diagnostic performance.[11,12] In the first study, pediatric neurologists were asked to read case vignettes and generate a ranked list of DDX and baseline workups (e.g., diagnostic studies). The clinicians then used the tool, and again provided a list of DDX and workups. The authors found the use of the tool significantly reduced the number of missing diagnoses in the DDX (36% to 15%; P<0.0001) across all clinicians and increased the relevance of the diagnoses listed.[11] In their second paper, Segal and colleagues evaluated the use of SimulConsult by nonspecialists to diagnose pediatric rheumatologic diseases via case vignettes. Similar to the earlier study, when using the DDX generator, the nonspecialists demonstrated a significant reduction in missed diagnoses in the DDX, which fell from 28 percent unaided to 15 percent using the tool (p<0.0001).[12]

Three papers provide evidence that DDX generators modestly improve the diagnostic accuracy of clinicians.[13-15] Using test patient cases in an exam format, Martinez-Franco et al. (2018) compared the diagnostic accuracy of first-year family medicine residents randomized to the control group with those in the intervention group, which used a DDX generator, DXplain. This tool requires the user to enter patients' signs, symptoms, and laboratory tests. Using these data, the tool generates a list of possible diagnoses ranked from highest to lowest probability. The mean percent-correct score and standard deviation was 74.1 ± 9.4 for the control group and 82.4 ± 8.5 for the intervention group (p<0.001).[13] Kostopoulou et al. (2017) developed a prototype DDX generator integrated with a commercial EHR system for use in general practice and tested it using high-fidelity simulation. As soon as the clinician enters the reason for encounter (RfE), the system generates a list of diagnostic suggestions based on the patient's RfE, age, and sex, and groups them according to published incidence rates (i.e., common, uncommon, and rare diagnoses). At the time of the study, the prototype supported three RfEs: chest pain, abdominal pain, and shortness of breath. Using standardized patients simulating 12 cases (4 cases per RfE), 34 general practitioners established their baseline performance with half of the cases and then used the DDX tool with the other half. Diagnostic accuracy improved significantly when using the tool, going from 49.5 percent to 58.3 percent accuracy (p<0.003).[14] Chou et al. (2017) tested the effect of a VCDDSS on the diagnostic accuracy of medical students and dermatology residents in a dermatology clinic. In this pilot study, the students' diagnostic accuracy increased significantly, from 62.5 percent without the VCDDSS to 81.25 percent using the VCDDSS (p<0.01).[15]

1.1.3.1.2 Specific Diagnoses

In addition to the differential diagnosis generators, the search identified papers that describe the development and evaluation of CDS models that determine whether a specific disease is present.

Several papers described rule-based or logic-based CDS for diagnosis where the tool had been integrated into a real clinical setting. Niemi et al. (2009) developed an automated CDS tool to identify patients admitted to the hospital with pneumonia or heart failure (HF) in real time to aid in timely administration of treatment. The system continually monitors data from existing information systems such as the pharmacy information system, laboratory management system, and radiology management system, and applies rules for pneumonia and HF. When the patient accumulates enough points to be diagnosed with either HF or pneumonia, the system looks to see whether appropriate treatment has been provided (e.g., in the case of pneumonia, an antibiotic) within set time limits, and if it has not, the system generates an alert to the clinician and nursing unit. In the emergency department (ED), the sensitivity and specificity of the system to identify pneumonia was 89 percent and 86 percent, respectively, and in the inpatient setting it was 92 percent and 90 percent, respectively. For HF, the sensitivity was 94 percent and the specificity 90 percent. In addition, the system allowed the hospital to increase compliance with national quality indicators for both of these conditions.[16]

Deleger et al. (2013) developed and tested an automated appendicitis—inflammation of the appendix—risk categorization algorithm for pediatric patients with abdominal pain, based on content from the EHR, and found this system to be comparable to use of physician experts. Using retrospective data, the CDS tool had an average F-measure of 0.867, with a sensitivity (recall) of 0.869 and a positive predictive value (precision) of 0.863.[17] Kharbanda et al. (2016) developed and implemented an electronic CDS tool for pediatric patients with abdominal pain that included a standardized abdominal pain order set, a web-based risk stratification tool, and an ordering alert. Compared with in the pre-implementation period, the trend of computed tomography (CT) scan use during the implementation period decreased significantly each month (p=0.007), and showed a 54-percent relative decrease in CT use in the post-implementation period. The authors found that the decrease in CT use was not associated with the potential unintended consequences of decreased use of CT: significant changes to the rates of appendectomies or missed appendicitis cases.[18]

Chamberlain et al. (2016) developed a mobile smart phone application for screening patients for pulmonary disease and conducted preliminary testing of the algorithms in a clinic setting. The application uses an electronic stethoscope, a method of digitizing peak flow meter readings, and patient questionnaire to identify patients with asthma and chronic obstructive pulmonary disease. The classification algorithms were successful in identifying patients with asthma and chronic obstructive pulmonary disease from the general patient population, with an area under the receiver operating characteristic curve of 0.97. Of note, during the study, patient breath sounds were auscultated using the electronic stethoscope but were evaluated by a pulmonologist. The authors note that they have since been able to develop algorithms to automatically identify abnormal lung sounds, making this technology possible for use by non-pulmonologists and potentially even non-clinicians to assist with diagnosis.[19]

One study focused on a CDS tool patients could use to aid in the screening of skin lesions. Wolf et al. (2013) investigated the use of four readily available smart phone applications designed to evaluate photographs of skin lesions and provide feedback on the risk of malignancy. Clinical images of previously diagnosed skin lesions were submitted for evaluation through the applications. The application with the highest sensitivity (98.1%) sent images directly to a board-certified dermatologist for analysis—essentially tele-dermatology. The sensitivity of the other three applications ranged from 6.8 percent to 70.0 percent, and they relied on automated algorithms to analyze the images.[20]

More-advanced CDS tools leveraging AI and machine learning have generated excitement over the potential to better augment clinician intelligence and support decision making. A cohort of the papers in our review describe models based on AI techniques to screen for and diagnose specific disorders and diseases. A systematic review by Wagholikar et al. (2011) includes 220 reports of new decision models or evaluations of existing models. The authors generalized their findings and concluded that these techniques have growing popularity for simple classifications but have yet to achieve an acceptable degree of accuracy, particularly for complex medical problems.[21] Other studies of AI identified beyond this systematic review all show promise in identifying disease, although the research continues to be investigational in nature, with a lack of implementation and testing in real clinical settings.[17,22-28]

1.1.3.2 CDS To Assist With Diagnostic Study Interpretation

Several papers included in this review described investigational studies of CDS tools to assist with diagnostic study interpretation, including imaging studies, electrocardiograms (ECGs), and pathology. Although these CDS tools are proof-of-concept in nature, they demonstrate the potential to augment clinician diagnostic performance but not completely replace it.

1.1.3.2.1 Use in Imaging

Three papers identified through the search focused on techniques to assist with interpretation of imaging studies. All were investigational in nature, describing the development and validation of the models.[27,29,30] Herweh et al. (2016) compared the diagnostic performance of an automated machine-learning algorithm to detect acute stroke on CT scans using a standardized scoring method to the performance of stroke experts and novices using the algorithm. Although this study had a small sample size, the automated tool showed similar scoring results to that of experts and better performance than the novices.[29] Bien et al. (2018) used deep learning, a subset of machine learning, to model the complex relationships between images and their interpretations. The model was designed to detect general abnormalities and two specific diagnoses (anterior cruciate ligament [ACL] tears and meniscal tears) on knee magnetic resonance imaging (MRI). For general abnormalities, there was no difference between the performance of the model and the general radiologists. For ACL tear detection, the model was highly specific but not significantly different from the specificity achieved by the radiologists. Radiologists achieved a significantly higher sensitivity (p=0.002) in detecting ACL tears. For meniscal tears, the radiologists achieved significantly higher specificity compared with the model (p=0.003). The authors also found that providing the radiologists with the predictions from the model improved their quality of interpretation of the MRI studies.[27] Li et al. (2018) developed an AI tool to detect nasopharyngeal malignancies under endoscopic evaluation by oncologists. Results indicate that the tool was significantly better in its performance compared with oncological experts; the overall accuracy was 88.0 percent (95% CI, 86.1 to 89.6) versus 80.5 percent (95% CI, 77 to 84).[30]

1.1.3.2.2 ECG Interpretation

In the evaluation of cardiac health, 12-lead ECGs are accompanied by computer interpretations to assist the clinician with diagnoses. These interpretations have been shown to often be inaccurate, primarily because of noisy background signals that interfere with automated pattern recognition by the machine algorithms. However, four studies in this review evaluated ECG interpretations by automated systems, and all found that the systems were no better or worse than human performance alone.[31-34]

Hughes et al. (2017) sought to improve ED workflow and reduce physician interruptions generated by the need to rapidly read triage ECGs for patients with chest pain. The authors examined the accuracy of ECGs identified as normal by the computer with the hypothesis that these normal ECGs would not have clinically significant findings. The negative predictive value of the normal computer interpretations was 99 percent (95% CI, 97 to 99), indicating that there may be a group of ECGs for which rapid physician re-interpretation is not necessary, thereby reducing interruptions.[31]

Two studies tested the accuracy of the diagnoses generated by the automated systems compared with human interpretation. Given that nonexpert ECG readers are more likely to rely on automated system interpretation for diagnosis, Hakacova et al. (2012) compared the accuracy of two different rhythm analysis software products with the accuracy of nonexpert readers and found no significant difference in performance. The authors also looked at the accuracy of the software for ECGs for which the diagnosis by the nonexpert was incorrect, and found that only 28 percent+/-10 percent (system A) and 25 percent+/-10 percent (system B) of the automated diagnoses were correct.[33] Mawri et al. (2016) examined whether the use of automated ECG interpretation would affect time to treatment for patients with ST-elevation myocardial infarction. The authors found that the computer-interpreted ECGs failed to identify 30 percent of patients with ST-elevation myocardial infarction and found significant differences in two quality-of-care measures: immediate emergency physician interpretation led to faster catheterization laboratory activation time ($p<0.029$) and faster median door-to-balloon time ($p<0.001$).[34] A study by Cairns et al. (2017) tested a semi-automated system that attempts to overcome the accuracy issues of automated systems by leveraging the strengths of human performance (i.e., the ability to recognize patterns through noisy signals). The system integrates a rule-based computer algorithm with interactive questions and prompts for the clinician to generate multiple diagnostic possibilities. The use of this semi-automated system increased the number of correct interpretations, but the increase was not statistically significant.[32]

1.1.3.2.3 Use in Pathology

Two studies evaluated the use of AI to aid in the diagnostic work of pathologists.[35,36] Vandenberghe et al. (2017) developed and evaluated the use of deep learning, an AI method, to identify specific cancer cell types. For 71 breast tumor samples, they found that the use of this computer-aided diagnosis tool had a concordance rate of 83 percent with pathologist review. The pathologist re-reviewed the 12 samples that had discordance between the diagnoses of the pathologist and the computer-aided diagnosis tool, prompting modifications to 8 of the original diagnoses.[35] Xiong et al. (2018), also using deep learning, developed and tested an AI-assisted method for the automatic detection of mycobacterium tuberculosis. Results showed high sensitivity (97.9%) and moderate specificity (83.6%), with 2 false negatives and 17 false positive cases due to contaminants.[36]

1.1.3.3 CDS To Identify Patients at Risk for Diagnostic Errors

Three studies examined the use of CDS tools to identify patients who are at risk of having a diagnostic error.[35,37,38] The systems were all effective at identifying at-risk patients and allowed potential diagnostic errors, including missed or delayed diagnoses, to be prevented, while saving the clinicians time by reducing manual workloads and cognitive burden. As previously discussed, the study by Vandenberghe et al. (2017) used discordance between the diagnoses generated by the AI tool and the diagnoses by the pathologist to flag cases where there may be a high risk of diagnostic error.[35]

Koopman et al. (2015) developed a system to compare final radiology reports with final ED diagnoses to ensure that the ED identified and appropriately treated an abnormality on radiologic examination. A text analysis system first screens radiology reports to identify limb abnormalities, including fractures, dislocations, and foreign bodies. If the system identifies an abnormality, the diagnosis is reconciled with the ED diagnosis, as defined by International Classification of Diseases, 10th Revision (ICD-10) codes. If there is a discrepancy, the chart is flagged as a possible misdiagnosis, allowing immediate review and followup. Across the three settings in which the study took place, 274 of 2,018 patients (13.6%) with radiologic abnormalities were flagged for potentially missed diagnoses, and the chart was reviewed manually. Nine of the cases were identified as truly missed diagnoses, and the other instances were due to the ED ICD-10 discharge diagnoses being ambiguous, and not indicative of a diagnostic error. The value in this method is that clinicians need to review only a small subset of the radiology reports, in this case 11 percent of the total number or radiology studies, to determine whether there were potentially missed diagnoses.[37]

Murphy et al. (2015) applied electronic triggers to EHR data to identify the presences of “red flags,” exclude records for which further evaluation is not warranted (e.g., patients in hospice), and identify the presence of a delay in diagnostic evaluation for three conditions: colon cancer, lung cancer, and prostate cancer. Examples of red flags include positive fecal occult blood testing for colon cancer, concerning imaging studies for lung cancer, and elevated prostate-specific antigen for prostate cancer. Delayed diagnostic evaluation was defined by the absence of documented followup action. The trigger flagged 1,256 patients out of 10,673 patients with abnormal findings (11.8%) as being high risk for delayed diagnostic evaluation. Of these, 749 were true positives, a positive predictive value of 59.6 percent. Times to diagnostic evaluation were significantly lower in intervention patients compared with control patients flagged by the colorectal trigger and prostate trigger. There was no significant difference for the lung trigger.[38]

1.1.3.4 Unintended Consequences

In general, the CDS tools have an added benefit of improving access to specialized care by providing the clinician with assistance in diagnosing conditions that would typically fall in the realm of a specialist.[12,19,27]

Several of the CDS tools identified in this review, in addition to improving diagnostic accuracy, would also allow prioritization of work, creating greater efficiencies and improving workflow once implemented in clinical settings.[27,31,38] These systems flagged studies or diagnoses that required followup, allowing the clinicians to prioritize their work.

For the CDS tools that generate DDX, Graber and Mathew (2008) raised the concern that presenting the clinician with a long list of diagnostic possibilities could be distracting or lead to unnecessary testing and procedures.[3] Elkin et al. (2010) suggested that these tools actually reduce the cost of care by assisting the clinician with a broader differential diagnosis list, which is more likely to contain the correct diagnosis. In the case of the DXplain tool, providing the list of diagnoses in order of likelihood can lead to the clinicians evaluating the more likely diagnoses earlier.[39]

1.1.4 Implementation

1.1.4.1 Facilitators

Since many of the studies were conducted to validate algorithms or were exploratory in nature (e.g., testing AI algorithms to determine their ability to predict correct diagnosis), few described experiences with implementation in real clinical settings.

In the meta-review of systematic reviews by Nurek et al. (2015), the authors determined the features and effectiveness of computerized diagnostic decision support systems for medical diagnosis in primary care. The authors identified conditions that need to be met if a fully integrated CDS tool for diagnoses is to be successfully implemented and used: the tool can readily be integrated into EHRs; is based on standard terminologies, such as diagnosis codes (e.g., ICD-10); has the ability to be easily updated; is thoughtfully integrated into the clinicians' cognitive workflow; and interfaces with the clinicians at appropriate action points.[40]

1.1.4.2 Barriers

The information generated by CDS for use in diagnosis is only as good as the information that is put into the system. For example, if the clinician interprets the physical exam incorrectly (e.g., saying that a physical sign is absent when it is present) and inputs that incorrect information into the tool, that error may negatively affect any diagnosis that is partially based on the presence of that sign.[10,11,25,37] In the study by Koopman et al. (2015), discharge diagnoses, as indicated by ICD-10 codes, are reconciled with the diagnosis from radiology reports. If the ICD-10 code is incorrect, the system may not recognize a potential missed diagnosis.[37] Gegundez-Fernandez et al. (2017) commented that accurate diagnosis can be achieved only if the clinician's assessment of the patients' signs and symptoms is correct, because the automated system will process only data that humans introduce.[10]

In the case of ECG interpretation, accurate ECG recording depends on many variables, including lead placement, weight, movement, coexisting electrolyte abnormalities, and symptoms. If the placement is wrong (e.g., leads are placed in wrong location), the interpretation may be wrong.[33,34]

1.1.5 Resources

Additional information can be found at the HealthIT.gov site, which offers information on how the use of EHRs can improve diagnosis (https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/improved-diagnostics-patient-outcomes) and through the National Academies report Improving Diagnosis in Health Care (http://www.nationalacademies.org/hmd/Reports/2015/Improving-Diagnosis-in-Healthcare).

1.1.6 Gaps and Future Directions

Although research in the use of CDS for diagnosis has been conducted for many years, there has been a failure to implement these tools widely, and published work continues to be predominantly that of exploratory studies in educational settings, testing of algorithms using retrospective data, or evaluation through simulation.[8] Wagholikar et al. (2012), in their systematic review of modeling techniques for diagnostic decision support, provided several suggestions for research and future work in this area, including evaluation of these applications in clinical settings.[21]

1.1.6.1 Leveraging the "CDS Five Rights" Approach

A useful framework for achieving success in CDS design, development, and implementation is the "CDS Five Rights" approach.[41] The CDS Five Rights model states that CDS-supported improvements in desired healthcare outcomes can be achieved if we communicate: (1) the right information: evidence-based, suitable to guide action, pertinent to the circumstance; (2) to the right person: considering all members of the care team, including clinicians, patients, and their caretakers; (3) in the right CDS intervention format, such as an alert, order set, or reference information to answer a clinical question; (4) through the right channel: for example, a clinical information system such as the EHR, a personal health record, or a more general channel such as the Internet or a mobile device; (5) at the right time in workflow, for example, at the time of decision/action/need. CDS has not reached its full potential in driving care transformation, in part because opportunities to optimize each of the five rights have not been fully explored and cultivated.[42]

Providing the Right Information to the End User: The process of integrating real-time analytics into clinical workflow represents a shift towards more agile and collaborative infrastructure building, expected to be a key feature of future health information technology strategies. As interoperability and big data analytics capabilities become increasingly central to crafting the healthcare information systems of the future, the need to address issues that ease the flow of health information and communication becomes even more important. Without tools that select, aggregate, and visualize relevant information among the vast display of information competing for visual processing, clinicians must rely on cues by “hunting and gathering” in the EHR. Alerts that embody “right information” should provide just enough data to drive end user action, but not so much as to cause overload.[43] Overload can create alert fatigue and lead to desensitization to the alerts, resulting in the failure to respond to warnings, both important and less important. Experience from the use of CDS in the medication ordering process has demonstrated this paradoxical increase in risk of harm due to alerts that were intended to improve safety.[44,45]

Providing Information in the Right Format: Lack of knowledge regarding how to present CDS to providers has impeded alert optimization, specifically the most effective ways to differentiate alerts, highlighting important pieces of information without adding noise, to create a universal standard. The potential solution that CDS represents is limited by problems associated with improper design, implementation, and local customization. In the absence of evidence-based guidelines specific to EHR alerting, effective alert design can be informed by several guidelines for design, implementation, and reengineering that help providers take the correct action at the correct time in response to recognition of the patient’s condition.[46]

Right Workflow: A well-thought-out user-centered design or equivalent process during the implementation phase includes critical elements of leadership buy-in, dissemination plans, and outcome measurements. Knowledge needs to be gained about how to implement the CDS and how to create an interface between the system and the clinician that takes into consideration the cognitive and clinical workflow.[27,47] The optimal approach to CDS should not be focused primarily—or even secondarily—on technology. Implementation is about people, processes, and technology. Systems engineering approaches, including consideration of user experience and improvements in user interface, can greatly improve the ability of CDS tools to reach their potential to improve quality of care and patient outcomes. The application of human factors engineering in determining the right workflow includes but is not limited to ethnographic research including workflow analysis and usability testing.

1.1.6.2 Trust in Automation

CDS is meant to augment clinician performance, not replace it, making it an imperative to carry existing work forward into actual clinical settings.[1] CDS has advanced to the point of becoming a “type of automation that supplements the human powers of observation and decision.” Technologies related to big data bring both exciting opportunities and worrying prospects for misinformation, disinformation, and falsified information. Further work is required to demonstrate clinical and economic evidence using data from a population representative of the health system in a way that clinicians find trustworthy.

1.1.6.3 Measurement

Successful CDS deployment requires evaluating not only whether the intended clinicians are using the tool at the point of care, but also whether CDS use translates into improvements in clinical outcomes, workflows, and provider and patient satisfaction. However, success measures are often not clearly enunciated at the outset when developing or implementing CDS tools. As a result, it is often difficult to quantify the extent to which CDS has been effectively deployed, as well as whether it is effective at managing the original diagnostic problem it was designed to address.

References for Section 1.1

1. El-Kareh R, Hasan O, Schiff GD. Use of health information technology to reduce diagnostic errors. BMJ Qual Saf. 2013;22(Suppl 2):ii40-ii51.
2. Coordinator OotN. Clinical Decision Support. https://www.healthit.gov/topic/safety/clinical-decision-support.
3. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med. 2008;23(1):37-40.
4. Schiff GD, Bates DW. Can electronic clinical documentation help prevent diagnostic errors? New Engl J Med. 2010;362(12):1066-9.
5. Bond WF, Schwartz LM, Weaver KR, et al.. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012;27(2):213-9.
6. Massalha S, Clarkin O, Thornhill R, et al. Decision support tools, systems, and artificial intelligence in cardiac imaging. Can J Cardiol. 2018;34(7):827-38.
7. Graber ML, Kissam S, Payne VL, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ Qual Saf. 2012;21(7):535-57.
8. Riches N, Panagioti M, Alam R, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. PloS One. 2016;11(3):e0148991.
9. David CV, Chira S, Eells SJ, et al. Diagnostic accuracy in patients admitted to hospitals with cellulitis. Dermatol Online J. 2011;17(3).
10. Gegundez-Fernandez JA, Fernandez-Vigo JI, Diaz-Valle D, et al. Uvemaster: a mobile app-based decision support system for the differential diagnosis of uveitis. Invest Ophthamol Vis Sci. 2017;58(10):3931-9.
11. Segal MM, Williams MS, Gropman AL, et al. Evidence-based decision support for neurological diagnosis reduces errors and unnecessary workup. Journal Child Neurol. 2014;29(4):487-92.
12. Segal MM, Athreya B, Son MBF, et al. Evidence-based decision support for pediatric rheumatology reduces diagnostic errors. Pediatr Rheumatol. 2016;14(1):67.
13. Martinez-Franco AI, Sanchez-Mendiola M, Mazon-Ramirez JJ, et al. Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial. Diagn. 2018;5(2):71-6.
14. Kostopoulou O, Porat T, Corrigan D, et al. Diagnostic accuracy of GPs when using an early-intervention decision support system: a high-fidelity simulation. Br J Gen Pract. 2017;67(656):e201-e8.
15. Chou W-Y, Tien P-T, Lin F-Y, et al. Application of visually based, computerised diagnostic decision support system in dermatological medical education: a pilot study. Postgrad Med J. 2017;93(1099):256-9.
16. Niemi K, Geary S, Quinn B, et al. Implementation and evaluation of electronic clinical decision support for compliance with pneumonia and heart failure quality indicators. Am J Health Syst Pharm. 2009;66(4):389-97.
17. Deleger L, Brodzinski H, Zhai H, et al. Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department.J Am Med Inform Assoc. 2013;20(e2):e212-e20.
18. Kharbanda AB, Madhok M, Krause E, et al. Implementation of electronic clinical decision support for pediatric appendicitis. Pediatr. 2016;137(5):e20151745.
19. Chamberlain DB, Kodgule R, Fletcher RR, et al. A mobile platform for automated screening of asthma and chronic obstructive pulmonary disease. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016: IEEE.

20. Wolf JA, Moreau JF, Akilov O, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. J Am Med Assoc Dermatol. 2013;149(4):422-6.
21. Wagholikar KB, Sundararajan V, Deshpande AW. Modeling paradigms for medical diagnostic decision support: a survey and future directions. J Med Syst. 2012;36(5):3029-49.
22. Arthi K, Tamilarasi A. Prediction of autistic disorder using neuro fuzzy system by applying ANN technique. Int J Dev Neurosci. 2008;26(7):699-704.
23. Lee Y-H, Hu PJ-H, Cheng T-H, et al. A preclustering-based ensemble learning technique for acute appendicitis diagnoses. Artificial intelligence in medicine. 2013;58(2):115-24.
24. Lin R-H. An intelligent model for liver disease diagnosis. Artif Intell Med. 2009;47(1):53-62.
25. Farmer N. An update and further testing of a knowledge-based diagnostic clinical decision support system for musculoskeletal disorders of the shoulder for use in a primary care setting. J Eval Clin Pract. 2014;20(5):589-95.
26. Song L, Hsu W, Xu J, et al. Using contextual learning to improve diagnostic accuracy: Application in breast cancer screening. IEEE J Biomed Health. 2016;20(3):902-14.
27. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. PLoS medicine. 2018;15(11):e1002699.
28. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc. 2016;316(22):2402-10.
29. Herweh C, Ringleb PA, Rauch G, et al. Performance of e-ASPECTS software in comparison to that of stroke physicians on assessing CT scans of acute ischemic stroke patients. Int J Stroke. 2016;11(4):438-45.
30. Li C, Jing B, Ke L, et al. Development and validation of an endoscopic images-based deep learning model for detection with nasopharyngeal malignancies. Cancer Commun. 2018;38(1):59.
31. Hughes KE, Lewis SM, Katz L, et al. Safety of Computer Interpretation of Normal Triage Electrocardiograms. Acad Emerg Med. 2017;24(1):120-4.
32. Cairns AW, Bond RR, Finlay DD, et al. A decision support system and rule-based algorithm to augment the human interpretation of the 12-lead electrocardiogram. J Electrocardiol. 2017;50(6):781-6.
33. Hakacova N, Trägårdh-Johansson E, Wagner GS, et al. Computer-based rhythm diagnosis and its possible influence on nonexpert electrocardiogram readers. J Electrocardiol. 2012;45(1):18-22.
34. Mawri S, Michaels A, Gibbs J, et al. The comparison of physician to computer interpreted electrocardiograms on ST-elevation myocardial infarction door-to-balloon times. Crit Pathw Cardiol. 2016;15(1):22-5.
35. Vandenberghe ME, Scott ML, Scorer PW, et al. Relevance of deep learning to facilitate the diagnosis of HER2 status in breast cancer. Sci Rep. 2017;7:45938.
36. Xiong Y, Ba X, Hou A, et al. Automatic detection of mycobacterium tuberculosis using artificial intelligence. Journal of thoracic disease. 2018;10(3):1936.
37. Koopman B, Zuccon G, Wagholikar A, et al., editors. Automated reconciliation of radiology reports and discharge summaries. AMIA Annual Symposium Proceedings; 2015: J Am Med Inform Assoc.
38. Murphy DR, Wu L, Thomas EJ, et al. Electronic trigger-based intervention to reduce delays in diagnostic evaluation for cancer: a cluster randomized controlled trial. J Clin Oncol. 2015;33(31):3560.
39. Elkin PL, Liebow M, Bauer BA, et al. The introduction of a diagnostic decision support system (DXplain™) into the workflow of a teaching hospital service can decrease the cost of service for

diagnostically challenging Diagnostic Related Groups (DRGs). Int J Med Inform. 2010;79(11):772-7.

40. Nurek M, Kostopoulou O, Delaney BC, et al. Reducing diagnostic errors in primary care. A systematic meta-review of computerized diagnostic decision support systems by the LINNEAUS collaboration on patient safety in primary care. Eur J Gen Pract. 2015;21(sup1):8-13.

41. Campbell RJ. The five rights of clinical decision support: CDS tools helpful for meeting meaningful use. J AHIMA. 2013;84(10):42-7 (web version updated February 2016).

42. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008;41(2):387-92.

43. Woods DD. The alarm problem and directed attention in dynamic fault management. Ergon. 1995;38(11):2371-93.

44. Slight SP, Seger DL, Nanji KC, et al. Are we heeding the warning signs? Examining providers' overrides of computerized drug-drug interaction alerts in primary care. PloS One. 2013;8(12):e85071.

45. Wright A, Aaron S, Seger DL, et al. Reduced effectiveness of interruptive drug-drug interaction alerts after conversion to a commercial electronic health record. J Gen Intern Med. 2018;33(11):1868-76.

46. Miller K, Mosby D, Capan M, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc. 2017;25(5):585-92.

47. Noirhomme Q, Brecheisen R, Lesenfants D, et al. "Look at my classifier's result": Disentangling unresponsive from (minimally) conscious patients. Neuroimage. 2017;145:288-303.

1.2 Patient Safety Practice: Result Notification Systems

Authors: Kendall K. Hall, M.D., M.S., and Gordon Schiff, M.D.

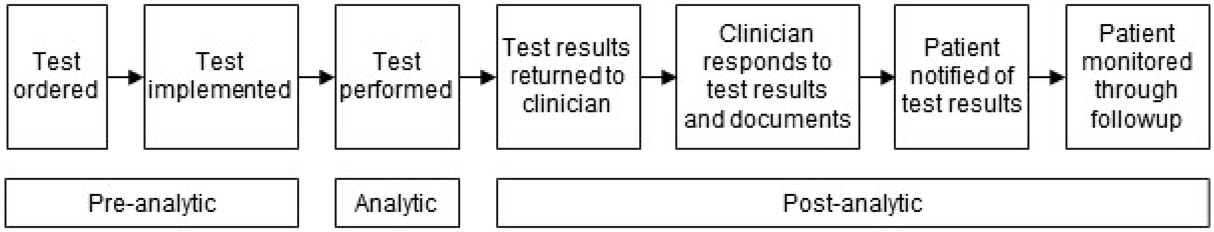
Reviewer: Andrea Hassol, M.S.P.H.

1.2.1 Practice Description

Failure to communicate test results has been repeatedly noted as a contributing factor to delayed diagnosis and treatment of patients in both ambulatory and inpatient settings.[1,2] Due to the negative impact on patients of missed communication of results, The Joint Commission made timely reporting of critical results of tests and diagnostic procedures a National Patient Safety Goal (NPSG.02.03.01) for their Critical Access Hospital and Hospital Programs.[3]

The laboratory and radiographic testing process has three distinct phases: the pre-analytic phase, during which the test is ordered and that order is implemented; the analytic phase, when the test is performed; and the post-analytic phase, in which results are relayed to the ordering clinician, who acts upon the results, and notifies and follows up with the patient (Figure 1).[4]

Figure 1: Conceptual Framework of the Testing Process[4]


Test ordered
Test implemented
Test performed
Test results returned to clinician
Clinician responds to test results and documents
Patient notified of test results
Patient monitored through followup
Pre-analytic
Analytic
Post-analytic

The post-analytic phase, specifically the step where results, clinically significant test results (CSTR) in particular, are relayed back to the ordering clinician, is a source of diagnostic error.[4,5] To reduce errors that occur during this step, experts have advocated for the use of automated alert notification systems to ensure timely communication of CSTR.[5-7] RNSs, which are the focus of this review, vary. They can be completely automated, where an abnormal result generates an alert to the ordering clinician; or the RNS may require manual activation by the clinician. There are also a variety of modalities that can be used to alert the practitioner of actionable test results, including short messages relayed via mobile phones; emails; and results (with or without accompanying alerts) in the EHR.[8]

1.2.2 Methods

The question of interest for this review is, "Do RNSs for radiologic and laboratory tests improve timeliness and reliability of receipt of results and action on the results?" To answer this question, we searched two databases (CINAHL® and MEDLINE®) for articles published from 2008 to 2018 using the terms "diagnostic errors," "delayed diagnosis," "missed diagnosis," and synonyms. Additional terms included "alerts," "automated systems," "communication systems," "critical test results," "alert notification," and other similar terms. The initial search yielded 1,965 results. Once duplicates had been removed and additional relevant articles from selected other sources added, a total of 1,981 articles were screened for inclusion, and 46 full-text articles were retrieved. Of those, 17 were selected for inclusion in this review, including 2 systematic reviews. Articles were excluded if the outcomes were not relevant to this review, the article was out of scope (including not quantitative), the study was of limited rigor, or if the study design or results were insufficiently described.

General methods for this report are described in the Methods section of the full report.

For this patient safety practice, a PRISMA flow diagram and evidence table, along with literature-search strategy and search-term details, are included in the report appendixes A through C.

1.2.3 Evidence Summary

The papers selected use RNS for CSTR, both life-threatening and nonurgent, for laboratory or radiological studies in inpatient and ambulatory settings. The RNS varied across studies and included both manual and automated mechanisms to generate the alert, and a variety of asynchronous and synchronous modalities to receive the alert. Outcomes included alerts being received (and acknowledged) by a clinician, and alerts being received and/or acted upon by the clinician (Table 1).

We reviewed one meta-analysis and one systematic review, both focusing on automated RNSs for laboratory results.[8,9] There were also several single studies of high-quality design, with two randomized controlled trials[10,11] and three cluster-randomized controlled trials.[12-14] Most of the single studies were quasi-experimental, with either pre/post or post-only designs.

1.2.3.1 Use of RNS for Radiologic Studies

Five studies focused on the impact of RNS on the communication of CSTR in radiology. The CSTR ranged from results requiring treatment but not immediately life-threatening to immediately life-threatening results. The impact of the RNS on the communication of results, and action taken on the results, was mixed.

Key Findings:

- Performance of result notification systems varied by type of test result, setting, synchronous versus asynchronous communication, and manual versus automated alerting mechanisms.
- For both critical and non-critical CSTR of radiologic studies, lab studies and tests pending at discharge, the use of RNS showed some positive but often mixed results in the timeliness and reliability of receipt, action acknowledgment, and action on the test results.
- Policies and procedures that aligned with the system, mindful integration of the RNS into the workflow and the EHR, and appropriate staffing were identified as factors supporting successful RNS.
- Significant barriers to successful implementation include poor system design, the lack of connectivity between hospitals and non-network physicians, challenges associated with changing schedules and providing critical alerts to physicians who may not be available, and variations in clinician response to alerted results.

Two studies, both by Lacson and colleagues, evaluated the use of an Alert Notification of Critical Results (ANCR) system to facilitate communication of critical imaging test results to ordering clinicians at a large academic medical center. The ANCR system, integrated into the clinical workflow, allows both synchronous communication (e.g., pagers) for results related to life-threatening conditions, and asynchronous communications (e.g., email). The system relies on radiologists who read and interpret the radiographic images to initiate an alert to the ordering clinician, rather than using a completely automated system. In the first study, the authors evaluated the ANCR system on adherence to a hospital policy for timeliness of notifications that is based on criticality of the imaging result.[15] Using a pre/post study design, the authors found a significant improvement in adherence to the timeliness policy, with adherence increasing from 91.3 percent before the ANCR intervention to 95.0 percent after ($p<0.0001$). In the second study, also using a pre/post study design, the authors evaluated the impact of implementing both the ANCR system and the policy of communication of the critical imaging test result in reducing critical results that lacked documented communication (date, time, and name of ordering clinician contacted). After the implementation of the critical imaging test result policy and the ANCR, critical results lacking documented communication decreased nearly fourfold between 2009 and 2014 (0.19 to 0.05, p<0.0001).[16]

Table 1: Overview of Single Studies

Author, Year	Clinically Significant Test Result Type & Severity	Result Notification System	Setting
Chen et al., 2011[21]	Laboratory—critical	Automated phone alert using short message service (SMS)	Inpatient/academic medical center
Dalal et al., 2014[12]	Test pending at discharge (TPAD)	Automated email system	Inpatient and outpatient/academic medical center
Dalal et al., 2018[13]	TPAD	Automated email system	Inpatient and outpatient/academic medical center
Eisenberg et al., 2010[10]	Radiologic—nonurgent	Manual, Web-based electronic messaging system	Academic medical center/inpatient and outpatient
El-Kareh et al., 2012[27]	TPAD	Automated email system	Inpatient and outpatient/academic medical center
Etchells et al., 2010[10]	Laboratory—critical	Automated paging system	Inpatient/academic medical center
Etchells et al., 2011[11]	Laboratory—critical	Automated alerts via mobile phone or pager and link to clinical decision support for alert	Inpatient/academic medical center
Lacson et al., 2014[15]	Radiology—critical	Manually triggered alert via pager or email	Inpatient/academic medical center
Lacson et al., 2016[16]	Radiology—critical	Manually triggered alert via pager or email	Inpatient/academic medical center
Lin et al., 2014[23]	Laboratory—critical	Automated phone text-message alert	Outpatient/academic medical center
O'Connor et al., 2016[17]	Radiology—nonurgent	Manually triggered alert via pager or email/alert in electronic medical record (EMR)	Outpatient/academic medical center
O'Connor et al., 2018[24]	Laboratory—nonurgent	Manually triggered alert via pager or email	Outpatient/academic medical center-affiliated community hospital
Park et al., 2008[20]	Laboratory—critical	Automated phone alert using SMS and callback	Inpatient/academic medical center
Singh et al., 2009[18]	Radiology—critical	Automated EMR alert notification system	Outpatient/U.S. Veterans Affairs (VA) medical center
Singh et al., 2010[5]	Laboratory—noncritical	Automated EMR alert notification system	Outpatient/VA medical center

In a study linked to the work of Lacson and colleagues, O'Connor et al. (2015) integrated an ANCR with an EHR-based results management application and evaluated its adoption and impact on followup of actionable results by primary care providers (PCPs) in the outpatient setting. Prior to integration, PCPs used the EHR application to track and acknowledge results from laboratory studies. The integration of the two systems allowed the PCPs to receive and acknowledge the ANCR-generated non-urgent CSTR alerts in the EHR or through the ANCR system. During the 2 years after implementation, 15.5 percent of the ANCR alerts were acknowledged in the EHR (15.6% year 1, 15.4% year 2). In the post-intervention period, there was a significant difference (p=.03) between the proportion of alerts acted upon that were acknowledged in the EHR application (79%; 95% CI, 52 to 92) compared with the alerts acknowledged in the ANCR system (97%; 95% CI, 90 to 99).[17]

Singh et al. (2009) evaluated the impact of an EHR-based system to alert clinicians to critical imaging results in a multidisciplinary ambulatory clinic at a large Veterans Administration (VA) medical center and its five satellite clinics. The VA EHR has an embedded notification system for alerting clinicians to CSTR in a "View Alert" window. The system requires that the radiologist reading an image flag abnormal imaging results, and these alerts are then transmitted to the "View Alert" window. During the study period there were 1,196 abnormal imaging alerts generated (0.97% of all imaging studies), and 217 (18.1%) of these alerts remained unacknowledged (i.e., the ordering clinician did not click on and open the alert) after 2 weeks. Using logistic regression, variables associated with a lack of acknowledgement included physician assistants compared with attending physicians (odds ratio [OR]: 0.46; 95% CI, 0.22 to 0.98); resident physicians compared with attending physicians (OR: 5.58; 95% CI, 2.86 to 10.89); and dual communication (i.e., communication with two clinicians) compared with communication with a single clinician (OR: 2.02; 95% CI, 1.22 to 3.36). Notably, 92 alerts, both acknowledged (n=71) and unacknowledged (n=21), lacked followup at 4 weeks.[18]

Eisenberg et al. (2010) evaluated the use of a Web-based electronic messaging system to communicate non-urgent CSTRs and recommend followup to ordering clinicians. As in the system used in the studies by Lacson and colleagues, the alerts were initiated by radiologists responsible for interpreting images through a web-based application. The request is received by a facilitator, who is then responsible for conveying the results to the ordering clinician. Once the results have been conveyed, the facilitator sends a confirmation back to the radiologist to close the loop. The authors recognized that the study design was weak (post-only with a satisfaction survey). They authors found that 82.2 percent of the alerts were communicated to the ordering clinicians within a 48-hour window, as defined by the time the radiologist submits a communication request to the time the facilitator conveys the communication to the ordering physician. The authors also found that the day of week affected the outcome, with more alerts submitted by the radiologists Monday–Thursday before 3 p.m. communicated within 48 hours (93.7% +/- 2.4), compared with alerts generated on Thursday afternoon through Sunday (73.0% +/- 9.2). The authors incidentally noted that for one-third of communications in which additional imaging or followup had been recommended, the electronic medical record had no documentation that these services were actually performed.[19]

1.2.3.2 Use of RNS for Laboratory Studies

Nine of the included studies focused on the use of RNS for laboratory studies, including one meta-analysis and one systematic review. As was the case for the RNS for radiologic studies, the evaluated interventions varied across studies and included paging, email, text messages, and EHR alerts. Results of the RNS were mixed.

The meta-analysis and the systematic review examined the effectiveness of automated electronic RNS to alert ordering clinicians to CSTR, and found insufficient/inconclusive evidence for the use of these systems.[8,9] The systematic review by Liebow et al. included four studies, two of which were used to calculate a standardized effect size (ES).[10,20] Etchells et al. reported results of a randomized controlled trial evaluating an automated RNS that sends critical laboratory values directly from the laboratory information system to a pager carried by the ordering physician. The objective was to evaluate the effect of the system on physician response time, defined as the time from when the critical result is entered into the lab system to the time an order is written in response to the critical value, or the documented time of treatment (whichever is relevant). They found a 23-minute reduction in median response time, 16 minutes (interquartile range [IQR] 2–141) for the automated paging group, and 39.5 minutes (IQR 7–104.5) for the usual care group, but this difference was not statistically significant.[10] Park et al. used a pre/post design to test the impact of a short message service and callbacks for action on critical hyperkalemia results. Across all patients in both intensive care units and general wards, the median and interquartile ranges for the clinical response times, defined as the frequency of clinical responses divided by the number of critical value alerts during a given time period, were significantly reduced, going from 213.0 minutes in the intensive care unit and 476.0 minutes in the general wards to 74.5 minutes and 241 minutes, respectively ($p<.001$).[20] Using Cohen's d, Liebow and colleagues calculated a grand mean reduction of time to communicate critical results for these two studies (d=.42; 95% CI, 0.23 to 0.62), indicating that the time to report a randomly selected CSTR using the automated system will be shorter than with a randomly selected manually reported value 61.8 percent of the time. Liebow et al. gave an overall strength of evidence rating of "suggestive" for automated RNS.[8]

Liebow et al. also conducted a systematic review of five studies evaluating the use of centralized call centers that communicate critical CSTRs to the ordering clinician.[8] Four of the five studies whose primary outcome was percent of calls completed within a specified interval after results were available from the laboratory (either <30 min or <60 min) contained sufficient data to calculate a standardized ES. The results of a random-effects meta-analysis support the implementation of call centers (mean OR=22.2; 95% CI, 17.1 to 28.7). This translates to critical lab values being reported faster with the call system than results reported via usual means (e.g., call to unit by laboratory technologist) approximately 88.6 percent of the time. Liebow et al. consider the overall strength of evidence for call center systems to be "moderate."

A systematic review by Slovis et al. included 34 articles published through 2016, representing 40 years of research related to asynchronous automated electronic laboratory RNS.[9] Although a wide variety of systems were represented and the study designs and outcomes differed, the authors summarized that these systems can be successfully implemented and improve timeliness of result notification and action. On closer examination of the five most recent studies that were included in the review and also identified through our search, the findings neither fully supported nor opposed use of these systems.[10,11,20-22]

In the first of two randomized controlled trials by Etchells et al. (2010) and included in the systematic review by Slovis et al., an automated paging system to convey critical laboratory results was evaluated in an urban academic medical center.[10] As described above, although there was a 23-minute reduction in the median response time, this was not statistically significant. In their second study, Etchells et al. (2011) combined an automated RNS with CDS.[11] The alerts, sent via text to a smart phone or to a pager, contained information about the specific patient and the abnormal result, and offered a URL to a webpage with decision support for the specific alert. The primary outcome was the proportion of pre-defined potential actions that were completed in response to the alert. A secondary outcome was the number of adverse events, defined as worsening of the patient's condition or complications related to the treatment of the condition. The median proportion of potential clinical actions that were completed was 50 percent (IQR 33–75%) with the alerting RNS with CDS and 50 percent (IQR 33–100%) without it, a difference that was not statistically significant, Without the system, there were 111 adverse events (33%) within 48 hours following an alert and with the alerting system on, there were 67 adverse events

(42%); a 9 percent increase when using the alerting system that bordered on statistical significance (p=0.06).

In addition to the five studies included in Slovis et al, two additional studies about laboratory RNS were reviewed for this report. In the outpatient department of a large (2,500-bed) tertiary teaching hospital in Taiwan, Lin et al. studied the impact of a phone-based RNS on clinical outcomes of patients taking the anticoagulant warfarin.[23] Their RNS automatically generates and delivers text messages about critical lab CSTRs to providers' mobile phones 24 hours a day, 7 days a week. Using a pre/post study design, the investigators found no significant differences in warfarin-associated adverse events. The rate of major venous thromboembolism events was 1.6 percent for both the manual alert period and the test RNS period. The rate of major hemorrhage requiring an ED visit or hospital admission was 3.1 percent in the manual alert period and 4.2 percent in the RNS alert period (p=0.198). As with the findings in Etchells et al. (2010), the secondary outcome of timeliness of physician followup actions after receipt of an automated critical alert was not significantly improved (11.13 ± 7.65 days for manual alert period vs. 11.32 ± 8.17 days for phone-based RNS period; p=0.814).

Expanding on the work previously described by Lacson and O'Connor, O'Connor et al. (2018) examined the use of the ANCR for communication of non-urgent clinically significant pathology reports indicating new malignancies.[24] After a pathologist identifies the CSTR, the ordering physician is contacted via pager about critical or urgent results, and via pager or secure email for non-urgent results, and the CSTR is entered into the ANCR system. For results that the ordering physician does not acknowledge, the system sends reminders to the pathologist and the ordering physician. Acknowledgment of the CTSR within 15 days, the institutional policy for non-urgent CSTR, was documented for 98 of 107 cases (91.6%) before the RNS had been implemented, and for 89 of 103 (86.4%) after the RNS had been implemented, a difference that was not statistically significant. There was also no significant difference in median time to acknowledgment for new malignancies when comparing the pre-RNS period (7 days; IQR 3–11) and post-intervention period (6 days; IQR 2–10). In the post-RNS period, for CTSR using the ANCR, median time to acknowledgment was significantly shorter than when an ANCR alert was not generated (2 vs. 7 days, p=0.0351).

1.2.3.3 Use of RNS for Tests Pending at Discharge

Tests pending at discharge (TPADs) involve transitions, span more than one setting (e.g., the hospital setting to the ambulatory setting), and often involve more than one clinician (e.g., inpatient attending physician and outpatient primary care physician). The risk of missed communication and potential harm to patients is greater during these transitions between settings and clinicians.[25,26] Three cluster-randomized controlled studies from a single institution investigated the use of an automated email CSTR notification system for TPAD.[12-14] Awareness and confirmed acknowledgement of the test result after discharge were statistically higher in the intervention group, but there was no difference between the intervention and control groups in documented actions taken in response to the test results (i.e., receiving/confirming receipt of a test result did not improve timeliness of acting upon that information).

Two cluster-randomized controlled studies by Dalal et al. (2014 and 2018) included inpatient attending physicians and PCPs whose patients were discharged from inpatient cardiology and medicine units in a large academic medical center, and who had TPAD for both radiology and laboratory studies.[12,13] In these studies, a patient's discharge triggers a series of electronic events that updates the status of any remaining TPADs on a daily basis. As results for these pending tests are finalized, the responsible inpatient attending and outpatient PCP receive an automatic email containing the test result. The primary outcome of the first study was self-reported awareness of the TPAD result by the patient's inpatient attending physician.[12] There was a statistically significant increase in the awareness of TPAD results by attending physicians for patients assigned to the intervention compared with those assigned to usual care (76% vs. 38%, adjusted/clustered OR 6.30, 95% CI, 3.02 to 13.16, p<0.001). The second study was larger, and the primary outcome was the proportion of actionable TPADs with documented action in the EHR.[13] For the primary outcome of documentation of action, there was no significant difference between the intervention and usual care groups (60.7% vs. 56.3%). For those that had an action documented, the median days between result notification and documented action was significantly lower in the intervention group (9 days, CI, 6.2 to 11.8) compared with the usual care group (14 days, 95% CI, 10.2 to 17.8) (p=0.04).

In the third study, by El-Kareh et al., the automated RNS described previously was used to alert inpatient and outpatient physicians about positive cultures when the final lab result was returned after patient discharge and the patient was not adequately treated with antibiotics. The alerts included patient identifiers, names and contact information for the physicians involved in their care, the culture results, the discharge medication list, and patient allergy information. Twenty-eight percent of results in the intervention group and 13 percent in the control group met the primary outcome of documented followup (in outpatient chart) within 3 days of receipt of the post-discharge lab result, a statistically significant difference [adjusted OR 3.2, 95% CI, 1.3 to 8.4; p=0.01].[27]

1.2.3.4 Unintended Consequences

Study authors raised a hypothetical concern about alert fatigue, a potential unintended consequence of implementing alerting RNSs, but only one study measured a related outcome: overuse of the alerting system. Lacson et al. (2016) found that the proportion of reports without critical CSTR and using the ANCR was significantly less than when the ANCR was not used (0.09 vs. 0.20, p<0.002, $\chi 2$ test).[16] Etchells et al. (2010) noted that critical results, such as those from repeated troponin tests, were viewed as nuisances by receiving clinicians during a pilot of the system.[10] They also noted that because physician schedules were not fully automated, it was not possible to consistently route critical results to a responsible *and available* physician to take action. To compensate for this, physicians handed off "critical value pagers" so that the physician-on-call carried several pagers. Although this could reduce the number of missed alerts, it also created confusion when the on-call physician often could not discern which pager was alerting.

Unexpectedly, Singh et al. (2009) found that dual communication, a duplication intended to ensure that at least one physician received the alert, was associated with delayed followup. This finding was attributed to the lack of clarity about who was responsible for handling the alert.[18]

1.2.4 Implementation

The studies included in this review demonstrated the critical importance of the local environment and technologies, and circumstances surrounding the success of RNS. Facilitators and barriers to implementation of RNSs are described below.

1.2.4.1 Facilitators

1.2.4.1.1 Integration of RNS Into Workflow

Dalal et al. (2014) attributed the successful implementation of their TPAD email-generating RNS to the existing institutional culture that supports the use of email as a routine part of clinical care. The RNS was integrated into their current practice, which facilitated uptake.[12] Two additional studies mention that alignment of the RNS with existing workflows minimized the need to actively seek results, and policies and procedures of the institution supported success.[17,27]

1.2.4.1.2 Clear Policies and Procedures for RNS Use

Several authors mentioned the need for clear policies and procedures for the RNS. Singh et al. (2009) indicated that institutions need to have clear policies about who is responsible for acknowledging an alert and taking action, so that there is no ambiguity.[18] One institution, after much deliberation, established the policy that the responsibility for following up a test rested on the "ordering" clinician, and that this responsibility could be discharged only after a handoff where the "new owner" recipient acknowledged receipt and agreed to take over the followup.[6] Other studies mentioned the need for policies establishing which types of alerts warrant use of the RNS and the timeliness of responding to those alerts, based on the criticality of the CSTR.[10,15-17,19]

1.2.4.1.3 Adequate Staffing To Support the RNS

Two studies mentioned the need for adequate staffing to support the implementation of RNS.[19, 21] Chen et al. implemented a two-pronged approach to improved communication times, involving increasing the number of staff in the laboratory to improve lab performance and quality, and implementing an RNS with secure messaging.[21] Eisenberg et al. noted that their RNS required the hiring of two full-time staff to manage the electronic messaging system.[19]

1.2.4.2 Barriers

1.2.4.2.1 Unaligned Policies and Procedures

Etchells et al. (2011) found that during weekends and nights there were differences in process between the study sites that involved the receipt of the alerts.[11] At one site, the smart phone on which alerts were received was handed from the attending daytime physician to the physician-on-call, so critical alerts could be received after hours. At the other site, the smart phone was not handed off, and the physician-on-call relied on telephone calls from the lab. O'Connor et al. (2018) documented that there were conflicting policies about what could trigger an alert: per local departmental policy, only unexpected malignancies should trigger an alert; but per enterprise policy, any new malignancy should trigger an alert.[24]

1.2.4.2.2 Lack of Connectivity Between Hospitals and PCPs Outside of Network

The three studies of using RNS to facilitate communication of TPADs during care transitions at hospital discharge all showed challenges in communicating with PCPs outside their hospital system.[12,13,27] If RNSs are relied on for TPAD result communication, they must be able to notify non-network and network PCPs.

1.2.4.2.3 Physician Handoffs and Scheduling

Automated physician scheduling is important for optimal performance of automated critical value alerting systems. This barrier to successful implementation was identified by Etchells et al. (2010), who found that when physician schedules are not fully automated, it is impossible to route alerts to the responsible (e.g., on-call) physician who can take action.[10,11]

1.2.4.2.4 Availability of Resources and Technology Limitations

Lin et al. (2014) indicated that the full implementation of their alert system was challenged by the unavailability of phones for adjunct physician staff, rendering them unable to receive critical alerts sent via the RNS.[23] Park et al. (2008) identified that their secure messaging phone reception had inconsistent signal strength in the hospital, but this had a minimal effect, since they had continued to manually call results to the unit in addition to the smartphone alerts.[20]

1.2.4.2.5 Financial Costs

There is an implied financial burden to implementing these systems, including costs of the systems themselves, and as mentioned previously, the potential need for increased staffing for successful implementation and use.[8,19,21]

1.2.5 Resources

The book "Getting Results: Reliably Communicating and Acting on Critical Test Results," (Schiff GD, ed., Joint Commission Resources, 2006) is "a collection of articles and case studies on how healthcare organizations are improving communication of critical test results," as described in the AHRQ Patient Safety Network.[28]

Pennsylvania passed the Patient Test Result Information Act (2018 Act 112) to ensure that patients with significant abnormalities on imaging exams are notified of the need for medical followup. Information on the law is available through the Pennsylvania General Assembly website: https://www.legis.state.pa.us/cfdocs/legis/li/uconsCheck.cfm?yr=2018&sessInd=0&act=112.]

1.2.6 Gaps and Future Directions

Over half of the studies in this review address the experiences of a small group of researchers from a single large academic institution and its affiliated medical centers. Although these studies are of high quality and some findings are significant, studies in other settings are needed to test and demonstrate generalizability, as well as to engage research in this field more widely. Diagnostic errors due to lapses in communication occur during care transitions, but only three studies (again, all in the same healthcare system) evaluated RNS to improve delivery of results finalized after the transition from the inpatient to the outpatient setting.

As mentioned above, it is challenging when many providers are taking care of a patient, as the RNS needs to discern who is responsible for which patient at any given time. Institutions are establishing policies aimed at addressing this challenge, but how the policies perform needs to be investigated.[6]

Another area for future study is the development and testing of RNS that are "smart" and use CDS to recognize the difference between critical results that require notification for emergent intervention versus those that do not. Future studies that track the number and types of alerts generated, including synchronous communication for only those CSTR that require urgent action, could include outcomes related to reducing alert fatigue.

References for Section 1.2

1. Callen JL, Westbrook JI, Georgiou A, et al. Failure to follow-up test results for ambulatory patients: a systematic review. J Gen Intern Med. 2012;27(10):1334-48.
2. Davis Giardina T, King BJ, Ignaczak AP, et al. Root cause analysis reports help identify common factors in delayed diagnosis and treatment of outpatients. Health Aff. 2013;32(8):1368-75.
3. Commission TJ. 2019 Hospital National Patient Safety Goals. In: Commission TJ, editor. 2019.
4. Hickner J, Graham D, Elder N, et al. Testing process errors and their harms and consequences reported from family medicine practices: a study of the American Academy of Family Physicians National Research Network. BMJ Qual Saf. 2008;17(3):194-200.
5. Singh H, Wilson L, Reis B, et al. Ten strategies to improve management of abnormal test result alerts in the electronic health record. J Patient Saf. 2010;6(2):121.
6. Roy CL, Rothschild JM, Dighe AS, et al. An initiative to improve the management of clinically significant test results in a large health care network. Jt Comm J Qual Patient Saf. 2013;39(11):517-27.
7. Singh H, Vij MS. Eight recommendations for policies for communicating abnormal test results. Jt Comm J Qual Patient Saf. 2010;36(5):226-AP2.
8. Liebow EB, Derzon JH, Fontanesi J, et al. Effectiveness of automated notification and customer service call centers for timely and accurate reporting of critical values: a laboratory medicine best practices systematic review and meta-analysis. Clin Biochem. 2012;45(13-14):979-87.
9. Slovis BH, Nahass TA, Salmasian H, et al. Asynchronous automated electronic laboratory result notifications: a systematic review. J Am Med Inform Assoc. 2017;24(6):1173-83.
10. Etchells E, Adhikari NK, Cheung C, et al. Real-time clinical alerting: effect of an automated paging system on response time to critical laboratory values—a randomised controlled trial. BMJ Qual Saf. 2010;19(2):99-102.
11. Etchells E, Adhikari NK, Wu R, et al. Real-time automated paging and decision support for critical laboratory abnormalities. BMJ Qual Saf. 2011;20(11):924-30.
12. Dalal AK, Roy CL, Poon EG, et al. Impact of an automated email notification system for results of tests pending at discharge: a cluster-randomized controlled trial. J Am Med Assoc. 2014;21(3):473-80.
13. Dalal AK, Schaffer A, Gershanik EF, et al. The impact of automated notification on follow-up of actionable tests pending at discharge: a cluster-randomized controlled trial. J Gen Intern Med. 2018;33(7):1043-51.
14. El-Kareh R, Hasan O, Schiff GD. Use of health information technology to reduce diagnostic errors. BMJ Qual Saf. 2013;22(Suppl 2):ii40-ii51.
15. Lacson R, Prevedello LM, Andriole KP, et al. Four-year impact of an alert notification system on closed-loop communication of critical test results. Am J Roentgenol. 2014;203(5):933-8.
16. Lacson R, O’Connor SD, Sahni VA, et al. Impact of an electronic alert notification system embedded in radiologists’ workflow on closed-loop communication of critical results: a time series analysis. BMJ Qual Saf. 2016;25(7):518-24.
17. O’Connor SD, Dalal AK, Sahni VA, et al. Does integrating nonurgent, clinically significant radiology alerts within the electronic health record impact closed-loop communication and follow-up? J Am Med Inform Assoc. 2016;23(2):333-8.
18. Singh H, Thomas EJ, Mani S, et al. Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: are electronic medical records achieving their potential? Arch Intern Med. 2009;169(17):1578-86.
19. Eisenberg RL, Yamada K, Yam CS, et al. Electronic messaging system for communicating important, but nonemergent, abnormal imaging results. Radiol. 2010;257(3):724-31.

20. Park H-i, Min W-K, Lee W, et al. Evaluating the short message service alerting system for critical value notification via PDA telephones. Ann Clin Lab Sci. 2008;38(2):149-56.
21. Chen T-C, Lin W-R, Lu P-L, et al. Computer laboratory notification system via short message service to reduce health care delays in management of tuberculosis in Taiwan. Am J Infect Control. 2011;39(5):426-30.
22. Singh H, Thomas EJ, Sittig DF, et al. Notification of abnormal lab test results in an electronic medical record: do any safety concerns remain? Am J Med. 2010;123(3):238-44.
23. Lin S-W, Kang W-Y, Lin D-T, et al. Comparison of warfarin therapy clinical outcomes following implementation of an automated mobile phone-based critical laboratory value text alert system. BMC Med Genomics. 2014;7(1):S13.
24. O'Connor SD, Khorasani R, Pochebit SM, et al. Semiautomated System for Nonurgent, Clinically Significant Pathology Results. Appl Clin Inform. 2018;9(02):411-21.
25. Roy CL, Poon EG, Karson AS, et al. Patient safety concerns arising from test results that return after hospital discharge. Annals of internal medicine. 2005;143(2):121-8.
26. Gandhi TK. Fumbled handoffs: one dropped ball after another. Ann Intern Med. 2005;142(5):352-8.
27. El-Kareh R, Roy C, Williams DH, et al. Impact of automated alerts on follow-up of post-discharge microbiology results: a cluster randomized controlled trial. J Gen Intern Med. 2012;27(10):1243-50.
28. AHRQ Patient Safety Network. Getting Results: Reliably Communicating and Acting On Critical Test Results. Book/Report. Agency for Healthcare Research and Quality. Accessed Feb. 18, 2020. https://psnet.ahrq.gov/issue/getting-results-reliably-communicating-and-acting-critical-test-results.

1.3 Patient Safety Practice: Education and Training

Authors: Kendall K. Hall, M.D., M.S., and Gordon Schiff, M.D.

Reviewer: Katharine Witgert, M.P.H.

1.3.1 Practice Description

In the 2015 National Academies of Sciences, Engineering, and Medicine (NASEM) report Improving Diagnosis in Health Care, one of the recommended strategies for improving diagnosis is to enhance healthcare professional education and training in the diagnostic process.[1] The content of this education can be guided by an understanding of the root causes of diagnostic errors. Studies have uncovered two broad categories of underlying root causes: cognitive-based factors, such as failed heuristics; and systems-based factors, such as lack of provider-to-provider communication and coordination.[2-4] In the realm of cognitive-based errors, there are also two main streams of thought about causes: heuristics failures and shortcomings in disease-specific knowledge and experience. These sets of broad conceptual factors are by no means mutually exclusive, and ideally system redesign and educational efforts can leverage overlaps and synergies. How to best provide education and training to change these underlying factors and thereby improve diagnostic accuracy and reduce diagnostic errors leads to a more fundamental question that this review attempts to address, “Do education and training lead to improved diagnostic performance?”

1.3.2 Methods

We searched four databases (CINAHL®, MEDLINE Cochrane, and PsycINFO®) for articles published from 2008 to 2018 using the terms “diagnostic errors,” “delayed diagnosis,” “missed diagnosis,” and synonyms. Terms specific to this PSP include “education, professional,” “training,” “simulation training,” “structured practice,” and related terms. The initial search yielded 211 results. Once duplicates had been removed and additional relevant referenced articles added, 187 articles were screened for inclusion and 29 full-text articles were retrieved. Of those, 22 studies were selected for inclusion in this review. Articles were included if the intervention being tested was training and an outcome was diagnostic accuracy. Articles were excluded if the article was out of scope, or the study provided limited detail or was of limited rigor.

Key Findings:

- Although there are a limited number of studies, the literature suggests that training on metacognitive skills may improve diagnostic accuracy, particularly as clinical experience increases.
- Online training, either didactic or via simulation, can be successfully used as a mode of delivery for educational interventions targeting clinical reasoning and diagnostic safety.
- There are several promising training interventions to improve visual perception for radiology practice.
- Limitations include a dearth of studies that examine the transfer of learning from the educational setting into the clinical setting and actual patient care.

General methods for this report are described in the Methods section of the full report.

For this patient safety practice, a PRISMA flow diagram and evidence table, along with literature-search strategy and search-term details, are included in the report appendixes A through C.

1.3.3 Evidence Summary

A majority of the selected studies focused on training directed at the cognitive aspects of diagnostic errors, such as clinical reasoning and biases. Other studies focused on training in visual perception skills for radiologists and specific diagnostic skills. Few studies involved experienced clinicians, with medical students and residents being the predominant types of learners.

Overall, the quality of evidence was moderate, with some strong study designs, such as randomized controlled trials, but with low numbers of subjects, making generalization of findings challenging. The educational interventions varied in both their content and the mode by which the content was delivered, and in several cases the distinction between the testing of the content versus the testing of the mode of delivery was difficult to ascertain.

1.3.3.1 General Training in Clinical Reasoning

Clinical reasoning is the process by which clinicians collect data, process the information, and develop a problem representation, leading to the generation and testing of a hypothesis to eventually arrive at a diagnosis.[5,6]

Cook et al. (2010) conducted a meta-analysis and systematic review of the effects on training outcomes of using virtual patients, including the effects on clinical reasoning. The learners interact with a computer program that simulates real-life clinical scenarios to obtain a history, conduct a physical exam, and make diagnostic and treatment decisions. In comparing virtual patients to no intervention, the pooled ES for the five studies with an outcome of clinical reasoning was 0.80 (95% CI, 0.52 to 1.08). Pooled ESs for the outcomes of knowledge (N=11 studies) and other skills (N=9 studies) were also large. When comparing the use of virtual patients to noncomputer instruction (e.g., didactic instruction, standardize patients, routine clinical activities), the pooled ES for the outcome of clinical reasoning was -0.004 (95% CI, -0.30 to 0.29, N=10 studies), and it was also low for satisfaction, knowledge, and other skills. The main takeaway from this meta-analysis and review was that the use of virtual patients is associated with large positive effects on clinical reasoning and other learning outcomes when compared with no intervention and is associated with small effects in comparison with noncomputer instruction.[7]

Graber et al. (2012) in their systematic review identified papers that reported testing interventions aimed at reducing cognitive errors.[8] Three broad categories of interventions emerged: interventions to improve knowledge and experience, interventions to improve clinical reasoning, and interventions that provide cognitive support. Several papers examined the use of training in metacognitive skills to improve clinical reasoning, which is below. Wolpaw et al. (2009) studied the use of a learner-centered technique by third-year medical students to present clinical cases in a structured manner (SNAPPS: Summarize, Narrow, Analyze, Probe, Plan, Select). Although the authors did not assess whether the DDX were accurate, they found that students using the SNAPPS technique performed better on all outcomes, including analyzing possibilities of the DDX, expressing uncertainties, and obtaining clarification.[9]

1.3.3.2 Training in Metacognitive Skills To Reduce Biases

Cognitive biases can affect clinical reasoning and influence the diagnostic process, contributing to a large proportion of misdiagnoses.[6,8,10,11] Metacognition, the understanding, control, and monitoring of one's cognitive processes, can be used to gain better insight and counteract these biases.[12,13] Nine studies focused on techniques to enhance metacognitive skills, specifically training on the use of cognitive forcing strategies (CFS) and the use of reflection during the diagnostic process. The results of the studies are mixed, but overall suggest the use of training metacognitive strategies to improve diagnostic performance.

The use of CFS, a metacognitive strategy, is a technique to bring about self-monitoring of decision making and to force the consideration of alterative diagnoses.[13] Three studies (Sherbino et al., 2011, Sherbino et al., 2014, Smith and Slack,2015) provided medical students and residents with training on the use of CFS and measured its impact on diagnosis.[14-16] The results did not show any appreciable improvement in diagnostic accuracy. In a preliminary and followup study, Sherbino et al. employed a 90-minute, standardized, interactive, case-based teaching seminar on CFS for medical students during their emergency medicine rotation.[14,15] Neither study showed any improvement in diagnostic errors. In the first study, they found that fewer than half of the students could use the CFS to debias themselves, and that 2 weeks post-training the students' knowledge of debiasing was no longer present. In the second study, there was no difference in the diagnostic accuracy between the control and intervention groups. In the study by Smith and Slack (2015), family medicine residents participated in a debiasing workshop that included training on CFS. They found that the residents' ability to formulate an acceptable plan to mitigate the effect of cognitive biases significantly improved after the training (p=0.02), although the residents were not able to translate the plan into practice, as evidenced by no change in the outcomes of preceptor concurrence with the residents' diagnoses, residents' ability to recognize their risk of bias, and the preceptors' perception of an unrecognized bias in the residents' presentations. This study was limited in that CFS targets biases related to nonanalytic reasoning (so-called pattern recognition). Novice diagnosticians, such as medical students, may lack sufficient experience to employ nonanalytic reasoning, rendering these methods increasingly more useful as experience increases.[16]

In a frequently cited study, Mamede et al. (2010) investigated whether recent diagnostic experiences elicit availability bias (i.e., judging a diagnosis that comes to mind more readily as being correct), and then tested a simple instructional procedure to reduce that bias. The training consisted of a five-step procedure to induce structured reflection and improve diagnostic accuracy in first- and second-year internal medicine residents. The use of reflection did reduce availability bias and improved diagnostic accuracy.[17] Two additional studies by this group of investigators (Mamede 2012, Mamede 2014), furthered this work on the use of structured reflection as a tool to facilitate diagnosis. The first of these studies found that the use of structured reflection after providing an immediate diagnosis when practicing with clinical cases fostered the learning of clinical knowledge more effectively than providing an immediate diagnosis only, or generating an immediate diagnosis followed by a differential diagnosis.[18] In the second study, the use of this technique enhanced learning of the diagnosis practiced as well as its alternative diagnoses.[19]

Coderre et al. (2010) tested the effectiveness of questioning a medical student's initial hypothesis as a means to induce cognitive reflection. The authors found that the questioning of an initial correct diagnosis did not change the final diagnosis; the students tended to retain the initial diagnosis. If the student's initial diagnosis was incorrect, the questioning provided an opportunity for the students to recognize and react to their error, and correct their diagnosis.[20]

In a randomized controlled study, Nendaz et al. (2011) studied the impact of weekly in-person case-based clinical reasoning seminars incorporating diagnostic reflection, during which the students were prompted to reflect on their reasoning process and were provided feedback on each step of that process. They found no difference in the accuracy of the medical students' final diagnoses between intervention and control groups (74% vs. 63%), although the students in the intervention group were more likely to have mentioned the correct diagnosis somewhere on their working list of DDX under consideration (75% vs. 97%, p=0.02).[21]

Reilly et al. (2013) incorporated the promotion of reflection on past experiences where a cognitive bias led to a diagnostic error, as part of a longitudinal curriculum on cognitive bias and diagnostic errors for residents. Residents who completed the curriculum significantly improved their ability to recognize cognitive biases when compared with their baseline performance (p=0.002) and when compared with the control group (p<0.0001). The study was limited in that it did not evaluate the impact of the intervention on diagnostic accuracy.[22]

1.3.3.3 Training on the Use of Heuristics

Heuristics are decision strategies, or mental shortcuts, that allow fast processing of information to arrive at a decision or judgment. One type of heuristic is representativeness; the use of the degree to which an event is representative of other, similar events to assess the probability of an event occurring.[23,24] Although the literature around the use of heuristics in medicine tends to focus on the biases they introduce, there is a recognized potential for training with heuristics to achieve better diagnostic accuracy.[25,26]

Mohan et al. (2018) conducted a randomized controlled trial comparing two training interventions designed to improve the use of the representativeness heuristic to improve trauma triage by emergency physicians. The authors developed two serious video games to train in the use of the heuristic. The first was an adventure game, based on the theory of narrative engagement, and the second was a puzzle-based game, based on the theory of analogical reasoning, using comparisons to help train the learners on applying decision principles. Both games incorporated feedback on diagnostic errors and how they could be corrected. Results showed that both games had positive effects on trauma triage, whereas traditional medical education had none.[26]

1.3.3.4 Training To Improve Visual Perception Skills

In radiology, diagnostic errors fall into two broad categories: perceptual errors, in which an abnormality on an image is not seen or identified, and interpretive errors, in which an abnormality is seen but the meaning or the importance of the finding is not correctly understood.[27,28] Perceptual errors account for a majority of misdiagnoses in radiology,[27,29,30] and can be rooted in faulty visual processing or, to a lesser extent, cognitive biases.[31]

Improving visual perception skills, which predominate the diagnostic process in radiology, requires methods of training different from those to improve clinical reasoning.[28] Four studies were identified through our search that evaluated the impact of educational interventions on perceptive skills, with three showing improvement in perceptive performance.[32-34] The studies involved subjects early in their medical training, and each tested a different intervention to improve perceptive performance, making aggregation of findings challenging.

A novel study by Goodman and Kelleher (2017) took 15 first-year radiology residents to an art gallery, where experts with experience in teaching fine art perception trained the residents on how to thoroughly analyze a painting. The trainees were instructed to write down everything they could see in the painting, after which the art instructor showed the trainees how to identify additional items in the painting that they had not perceived. To test this intervention, the residents were given 15 radiographs pre-intervention and another 15 post-intervention and asked to identify the abnormalities. At baseline, the residents scored an average of 2.3 out of a maximum score of 15 (standard deviation [SD] 1.4, range 0–4). After the art training, the residents' scores significantly improved, with an average score of 6.3 (SD of 1.8, range 3–9, p<.0001), indicating that perception training may improve radiology residents' abilities to identify abnormalities in radiographs.[32]

In a small randomized crossover study by van der Gijp et al. (2017), 19 first- and second-year radiology residents received training on two different visual search strategies to determine their effect on accuracy of detecting lung nodules on CT scans. The first search strategy was "scanning," in which the resident views all the visible lung tissue on a single image and slowly scrolls down, image by image, through the entire study. The second search strategy, "drilling," has the resident mentally divide each lung into three regions and scroll through each region individually while keeping the eyes fixed on that region. Perceptual performance for both scanning and drilling strategies and a pre-test using a free search strategy was determined by the mean numbers of true positives and false positives. There was a significant effect of year of residency on the true positive score (p<0.01) but not for false positives. Drilling (p<0.001) and free search (p<0.001) both resulted in significantly higher true positive scores (i.e., lung nodules identified appropriately) than did scanning. There was no difference between the drilling or free search strategies. The free search strategy resulted in higher false positive scores than drilling (p<0.001) and scanning (p<0.001). The authors concluded that drilling outperforms scanning for detecting lung-nodules and should be the preferred strategy when teaching perceptive skills.[33]

In a randomized controlled trial, Soh et al. (2013) used an online e-learning tutorial to train 14 first-year medical radiation sciences students (i.e., radiology technologists) in Australia to improve their ability to detect breast lesions on mammographic images. The 1-hour tutorial focused on anatomy, image positioning, mammogram viewing and analysis, and the appearance of normal breast tissue and asymmetric densities and masses. The students were randomized to the intervention (tutorial) or control group, with their performance evaluated by their viewing normal and abnormal mammograms. The study used eye-tracking technology to determine when and how often the student fixated on a lesion. The intervention group demonstrated improvement in the mean number of fixations per case (p=.047), and decreased time to first fixation on a lesion by 49 percent (p=.016). The intervention increased students' ability to identify lesions (i.e., sensitivity) by 30 percent (p=.022).[34]

The fourth study evaluated different proportions of normal and abnormal radiographs in image training sets to determine the best case-mix for achieving higher perceptive performance.[35] For the intervention, Pusic et al. (2012) used three different 50-case training sets, which varied in their proportions of abnormal cases (30%, 50%, 70%). One hundred emergency medicine residents, pediatric residents, and pediatric emergency medicine fellows were randomized to use one of the training sets. After the intervention, all participants completed the same post-test. All three groups showed improvement after the intervention, but with varying sensitivity-specificity trade-offs. The group that received the lowest proportion (30%) of abnormal radiographs had a higher specificity and was more accurate with negative radiographs. The group that trained on the set with the highest proportion of abnormal radiographs (70%) detected more abnormalities when abnormalities were present, achieving higher sensitivity. These findings have significant implications for medical education, as it may be that case mix should be adjusted based on the desired sensitivity or specificity for a given examination type (e.g., screening exams vs. diagnostic test).[35]

1.3.3.5 Other Education and Training Interventions

In the systematic review of patient safety strategies targeted at diagnostic errors, McDonald et al. (2013) identified 11 studies, ranging in dates from 1981 to 2011, that involved a variety of interventions.

Two randomized trials targeted patients and families, and found that the interventions improved performance: the first found that parent education improved parents' ability to identify serious symptoms requiring a physician office visit, and the second showed that patient education, in addition to reminders, improved breast cancer screening rates.[36]

Medical teaching typically involves error avoidance training (EAT), in which the focus is on how to perform a task correctly rather than on how to manage errors. However, evidence suggests that in the early stages of learning a skill, errors are necessary in order to avoid them in the future.[37] In their randomized trial with 56 medical students, Dyre et al. (2017) compared the use of error management training (EMT), in which students were given "permission" to make errors while conducting simulation-based ultrasound examinations, with EAT. Two outcomes were measured: the objective structured assessment of ultrasound scale, a measure of ultrasound performance; and diagnostic accuracy. The scale scores showed a significant improvement by the EMT group compared with the EAT group ($p<0.001$). Although diagnostic accuracy showed some improvement, this was not statistically significant. The study is limited in that the authors cannot determine whether the outcomes were a result of EMT, the positive framing of errors, or the combination of these two components.[38]

In a quasi-randomized controlled trial, Schwartz et al. (2010) tested the use of in-person didactic sessions to teach fourth-year medical students skills in contextualizing patient care. The authors based this work on the premise that there are both biomedical and contextual data that must be ascertained during the diagnostic process and incorporated into the treatment plan. Students who participated in the didactic sessions were significantly more likely to probe for contextual issues and significantly more likely to develop appropriate treatment plans for the standardized patients with contextual issues.[39]

Two studies investigated the use of online training to improve specific diagnostic skills, both resulting in significant improvements in diagnostic accuracy.[40,41] Smith et al. (2009) conducted a 4-month online didactic continuing education program to improve the ability of radiographers in rural areas to interpret plain musculoskeletal radiographs. The results showed a significant improvement in image interpretation accuracy for more complex cases ($p<0.05$), although there was no change in accuracy for less complex cases.[40] McFadden et al. (2016), using convenience sampling, compared a traditional in-person training with an on-line simulation-based training, both designed to improve the diagnosis by primary care practitioners of the etiology of joint pain. The online training included interactive practice opportunities and feedback delivered by an AI-driven tutor. The intervention group's diagnostic performance was significantly improved from baseline ($p<0.02$) compared with the group that received the traditional training.[41]

1.3.4 Implementation

Many of the studies were conducted in training and simulated environments. As such, there were limited discussions regarding facilitators and barriers to implementation in the clinical practice environment.

1.3.4.1 Facilitators

Several of the studies used interventions that brought the education to the learner, such as those using information technology–based platforms.[34,35,40,41] Although there are costs in both time and money associated with the set-up of these online learning systems, once implemented, the training is more easily administered to more learners and is more flexible to being customized.[42]

1.3.4.2 Barriers

The use of cognitive training interventions, such as reflective practice, may yield the greatest improvements for only the most complex diagnostic cases.[17,21] This makes application of appropriate strategies in actual clinical settings difficult, as whether a case is complex is often not determined until after the diagnostic process has begun.

In addition, some of these teaching techniques, such as those using standardized patients or requiring development of simulations, are labor intensive and may not be generalizable.

1.3.5 Resources

The Society to Improve Diagnosis in Medicine offers a clinical reasoning toolkit that contains resources to help clinicians and educators. The toolkit can be accessed at the Society's website: https://www.improvediagnosis.org/clinicalreasoning/.

1.3.6 Gaps and Future Directions

Graber et al. (2012) noted that the research directed at improving cognition as a means to reduce diagnostic errors was immature, and recommended more research to study different approaches and interventions, including training strategies.[8] Since the publication of that paper, as evidenced in this review, there has been a modest increase in research focused on education and training to improve diagnostic accuracy, but gaps persist and many questions remain unanswered and untested. There is a particularly strong need to be able to take the educational work out of the realm of the classroom and into the real and complex world that busy diagnosticians work in to reliably make accurate and timely diagnoses.

It should be noted that interventions, such as education and training, that address human error directly are considered to be weaker at driving effective and sustained improvement compared with stronger actions that remove dependence on the human.[43] There is an opportunity to identify and investigate the use of system supports (e.g., process changes) to complement and solidify these educational efforts. The supports should prevent the need for clinicians to rely solely on their human, hence fallible, memory to recall what they learned both about diseases and about heuristics pitfalls.

Many of the studies engaged medical students or first-year residents, who were relatively new in their careers and lacked clinical experience. Improving clinical reasoning through the use of metacognitive strategies, particularly CFS, is targeted at reducing biases associated with the use of nonanalytic reasoning. Expanding training on these strategies to more experienced clinicians, as opposed to trainees, may yield stronger results.[14,15] In contrast, visual perceptive skills develop earlier and faster than interpretive skills.[28,44] Therefore, educational interventions directed at improving perception would more likely benefit medical students and residents early in their training. Understanding the best timing for different educational strategies to maximize their effectiveness in the continuum of medical education, from student through experienced clinician, would be beneficial.

A variety of methodological aspects of the studies could have been improved to strengthen the evidence base. Several of the studies occurred outside of clinical settings (e.g., online training and testing), and did not involve transferring the skills to diagnostic accuracy outcomes in actual clinical practice. Some of the studies were limited due to the inability to untangle the effect of the mode of the training (e.g., online didactic training) from the content that was being delivered.

Finally, although research indicates that the root causes of diagnostic errors include both cognitive factors and systems-based factors,[2-4] nearly all the identified studies targeted cognitive-based training strategies. The 2015 NASEM report on improving diagnosis in healthcare included a call for training in systems-based factors. This is an opportunity to conduct studies on the impact of team-training and communication on diagnostic errors, which is lacking, and training to support patient integration into the diagnostic process.

References for Section 1.3

1. National Academies of Sciences, Engineering and Medicine. 2015. Improving Diagnosis in Health Care. Washington, DC: National Academies Press. https://doi.org/10.17226/21794.
2. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005;165(13):1493-9.
3. Kostopoulou O, Delaney BC, Munro CW. Diagnostic difficulty and error in primary care—a systematic review. Fam Pract. 2008;25(6):400-13.
4. Singh H, Giardina TD, Meyer AN, et al. Types and origins of diagnostic errors in primary care settings. J Am Med Assoc Intern Med. 2013;173(6):418-25.
5. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. New Engl J Med. 2006;355(21):2217-25.
6. Norman GR, Monteiro SD, Sherbino J, et al. The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Acad Med. 2017;92(1):23-30.
7. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med. 2010;85(10):1589-602.
8. Graber ML, Kissam S, Payne VL, et al. Cognitive interventions to reduce diagnostic error: a narrative review. BMJ Qual Saf. 2012;21(7):535-57.
9. Wolpaw T, Papp KK, Bordage G. Using SNAPPS to facilitate the expression of clinical reasoning and uncertainties: a randomized comparison group trial. Acad Med. 2009;84(4):517-24.
10. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Med Educ. 2010;44(1):94-100.
11. van den Berge K, Mamede S. Cognitive diagnostic error in internal medicine. Euro J Intern Med. 2013;24(6):525-9.
12. Colbert CY, Graham L, West C, et al. Teaching metacognitive skills: helping your physician trainees in the quest to'know what they don't know'. Am J Med. 2015;128(3):318.
13. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78(8):775-80.
14. Sherbino J, Dore KL, Siu E, et al. The effectiveness of cognitive forcing strategies to decrease diagnostic error: an exploratory study. Teach Learn Med. 2011;23(1):78-84.
15. Sherbino J, Kulasegaram K, Howey E, et al. Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning: a controlled trial. Can J Emerg Med. 2014;16(1):34-40.
16. Smith BW, Slack MB. The effect of cognitive debiasing training among family medicine residents. Diagn. 2015;2(2):117-21.
17. Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. J Am Med Assoc. 2010;304(11):1198-203.
18. Mamede S, van Gog T, Moura AS, et al. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. Med Educ. 2012;46(5):464-72.
19. Mamede S, Van Gog T, Sampaio AM, et al. How can students' diagnostic competence benefit most from practice with clinical cases? The effects of structured reflection on future diagnosis of the same and novel diseases. Acad Med. 2014;89(1):121-7.
20. Coderre S, Wright B, McLaughlin K. To think is good: querying an initial hypothesis reduces diagnostic error in medical students. Acad Med. 2010;85(7):1125-9.
21. Nendaz M, Gut A-M, Louis-Simonet M, et al. Bringing explicit insight into cognitive psychology features during clinical reasoning seminars: a prospective, controlled study. Educ Health. 2011;24(1):496.

22. Reilly JB, Ogdie AR, Von Feldt JM, et al. Teaching about how doctors think: a longitudinal curriculum in cognitive bias and diagnostic error for residents. BMJ Qual Saf. 2013;22(12):1044-50.
23. Kahneman D, Tversky A. Subjective probability: A judgment of representativeness. Cogn Psychol. 1972;3(3):430-54.
24. Kahneman D. Thinking, fast and slow: Macmillan; 2011.
25. Marewski JN, Gigerenzer G. Heuristic decision making in medicine. Dialogues Clin Neurosci. 2012;14(1):77.
26. Mohan D, Fischhoff B, Angus DC, et al. Serious games may improve physician heuristics in trauma triage. Proc Natl Acad Sci. 2018;115(37):9204-9
27. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics. 2015;35(6):1668-76.
28. Norman GR, Coblentz CL, Brooks L, et al. Expertise in visual diagnosis: a review of the literature. Acad Med. 1992;67(10):S78-83.
29. Berlin L. Radiologic errors, past, present and future. Diagn. 2014;1(1):79-84.
30. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. J Med Imag Radiat On. 2012;56(2):173-8.
31. van der Gijp A, Ravesloot C, Jarodzka H, et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. Adv Health Sci Educ. 2017;22(3):765-87.
32. Goodman TR, Kelleher M. Improving novice radiology trainees' perception using fine art. J Am Coll Radiol. 2017;14(10):1337-40.
33. van der Gijp A, Vincken KL, Boscardin C, et al. The effect of teaching search strategies on perceptual performance. Academic Radiol. 2017;24(6):762-7.
34. Soh BP, Reed WM, Poulos A, et al. E-tutorial improves students' ability to detect lesions. Radiologic technology. 2013;85(1):17-26.
35. Pusic MV, Andrews JS, Kessler DO, et al. Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. Med Educ. 2012;46(3):289-98.
36. McDonald KM, Matesic B, Contopoulos-Ioannidis DG, et al. Patient safety strategies targeted at diagnostic errors: a systematic review. Ann Intern Med. 2013;158(5_Part_2):381-9.
37. Rogers DA, Regehr G, MacDonald J. A role for error training in surgical technical skill instruction and evaluation. Am J Surg. 2002;183(3):242-5.
38. Dyre L, Tabor A, Ringsted C, et al. Imperfect practice makes perfect: error management training improves transfer of learning. Med Educ. 2017;51(2):196-206.
39. Schwartz A, Weiner SJ, Harris IB, et al. An educational intervention for contextualizing patient care and medical students' abilities to probe for contextual issues in simulated patients. J Am Med Assoc. 2010;304(11):1191-7.
40. Smith T, Traise P, Cook A. The influence of a continuing education program on the image interpretation accuracy of rural radiographers. Rural Remote Health. 2009;9(2).
41. McFadden P, Crim A. Comparison of the effectiveness of interactive didactic lecture versus online simulation-based CME programs directed at improving the diagnostic capabilities of primary care practitioners. J Contin Educ Health Prof. 2016;36(1):32-7.
42. O'Doherty D, Dromey M, Lougheed J, et al. Barriers and solutions to online learning in medical education–an integrative review. BMC Med Educ. 2018;18(1):130.
43. Foundation NPS. RCA2: improving root cause analyses and actions to prevent harm. National Patient Safety Foundation Boston, MA; 2015.

44. Brazeau-Lamontagne L, Charlin B, Gagnon R, et al. Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. Med Teach. 2004;26(4):326-32.

1.4 Patient Safety Practice: Peer Review

Authors: Kendall K. Hall, M.D., M.S., and Gordon Schiff, M.D.
Reviewer: Katharine Witgert, M.P.H.

1.4.1 Practice Description

Peer review is the systematic and critical evaluation of performance by colleagues with similar competencies using structured procedures.[1] Peer review in clinical settings has two recognized objectives: data collection and analysis to identify errors; and feedback with the intention of improving clinical performance and practice quality.[2,3] It also serves to fulfill accreditation requirements, such as The Joint Commission requirement that all physicians who have been granted privileges at an organization undergo evaluation of and collect data relating to their performance, or the American College of Radiology physician peer review requirements for accreditation.[4-6] When done systematically and fairly, peer review contributes to and derives from a culture of safety and learning.[7]

For this PSP, we are focusing the use of peer review as a tool to help identify, analyze, and discuss failures in establishing timely and accurate diagnoses, as well as a method to reduce diagnostic errors in the future.

Peer review, when designed appropriately, has the potential to achieve patient safety goals by having an impact on care either directly at the time of testing (e.g., identifying and resolving the error before it affects the patient) or indirectly by improving physician practice through continual learning and feedback. Thus, the question of interest for this review is, “Do peer review and feedback lead to improved diagnostic performance, i.e., fewer diagnostic errors?”

Key Findings:

- Second reviews of radiology or pathology interpretations by peers consistently uncover small but significant numbers of misread tests.
- The existence of any positive outcomes from increasing awareness of this general vulnerability and its effects on personal accountability—knowing that readings are being scrutinized—cannot be determined from the published studies.
- There is a lack of evidence to show that traditional random peer review and feedback mechanisms, which are used to maintain compliance with accreditation requirements, improve diagnostic quality over time or prevent diagnostic errors from reaching the patient.
- When nonrandom peer review is conducted prospectively, there is an opportunity to identify and remediate the diagnostic error before it reaches the patient.
- Limiting peer review to specific case types where it was most impactful was identified as a factor supporting implementation.
- Significant barriers to successful implementation include the increased staffing needs, workload, associated costs, concern over fairness, and maintenance of confidentiality of clinician performance.

1.4.2 Methods

We searched three databases (CINAHL®, MEDLINE, and PsycINFO®) for articles published from 2008 to 2018 using the terms “diagnostic errors,” “delayed diagnosis,” “missed diagnosis,” and synonyms. Terms specific to this PSP include “peer review,” “performance review,” “performance feedback,” “feedback,” “quality assurance,” and related terms. The initial search yielded 426 results. Once duplicates had been removed and additional relevant referenced articles added, 334 articles were screened for inclusion and 42 full-text articles were retrieved. Of those, 16 studies were selected for inclusion in this review. Articles were excluded if the focus was on the use of peer review in medical student or resident education, the outcome was not relevant to this review, the article was out of scope, or the study was of significantly limited rigor.

General methods for this report are described in the Methods section of the full report.

For this patient safety practice, a PRISMA flow diagram and evidence table, along with literature-search strategy and search-term details, are included in the report appendixes A through C.

1.4.3 Evidence Summary

A summary of key findings related to the use of peer review and feedback to reduce diagnostic error is located in the text box above. After the inclusion and exclusion criteria are applied, the preponderance of published literature about peer review related to diagnosis is from the specialties of radiology and pathology, likely due to findings that are "fixed" in images or specimens, leaving an artifact of the error that the review process can identify. It is also likely a testament to the leadership in these specialties, who have engaged their practitioners in responsibly reviewing their peers and caring for their patients.

Overall, the quality of evidence was moderate, with many descriptive study designs characterizing the rate and types of missed diagnoses using peer review. Studies also noted that scoring of radiological discrepancies is subjective and has significant variations in interrater reliability.[8]

The selected studies were categorized into two types of peer review, random and nonrandom, based on the methods of case selection.[5] Random peer review is characterized by the random selection of cases for review. There are several types of nonrandom review, including double reading of selected cases, review of diagnostically complex cases, and review of cases where potential diagnostic errors have been identified.

1.4.3.1 Traditional Peer Review: Random Versus Nonrandom Selection

Evaluation of professional practice, which can be accomplished through peer review, is a requirement for accreditation by organizations such as the American College of Radiology (ACR) and The Joint Commission, and recommended by professional associations such as the College of American Pathologists. The best-known example is that used in radiology, the ACR's RADPEER™ program, which is a standardized process with a set number of cases targeted for review (typically 5%) and a uniform scoring system.[9] The cases, which are originally interpreted images being used for comparison during a subsequent imaging exam by the reviewing "peer" radiologist, are randomly selected and scored.[5,6] Scores are assigned based on the clinical significance of the discrepancy between the initial radiologist's interpretation and the review radiologist's interpretation: (1) concur with interpretation; (2) discrepancy in interpretation, correct interpretation is not ordinarily expected to be made (i.e., an understandable miss); and (3) discrepancy in interpretation and the correct interpretation should be made most of the time. Scores of 2 and 3 can be modified with an additional designation of (a) unlikely to be clinically significant or (b) likely to be clinically significant. Scores of 2b, 3a, or 3b are reviewed by a third party, typically a department chair, medical director, or quality assurance committee.[9] Discrepancy rates can then be calculated for individual radiologists and used for comparison against peer groups or national benchmarks, and for improving practice.

Six studies involved the use of random peer review strategies similar to that of RADPEER. Each of these studies calculated discrepancy rates of case interpretation between the initial physician's diagnosis and the peer reviewer's diagnosis, with some studies using a third party to adjudicate the presence and severity of a diagnostic error.[10-13] Four of the studies compared random versus nonrandom approaches.[11-14] Table 2 lists the studies by case type, case selection, and discrepancy (i.e., diagnostic error) rates. The definition of discrepancy varied slightly across studies, but typically ranged from minor disagreements between reviewers that would necessitate a change to a report but are incidental to treatment, to major disagreements that may directly affect a patient's outcome.

Table 2: Discrepancy Rates for Peer Review in Pathology and Radiology

Author, Year	Case Type	Case Selection	Discrepancy Rates
Harvey et al., 2016[10]	Radiology	Random—consensus-oriented group review of random cases	2.7% (306/10,852)
Itri et al., 2016[11]	Radiology	Random—each radiologist reviews 20 random cases/month	2.6% (44/1,646)
		Nonrandom—submitted cases of potential diagnostic errors	100% (190/190)*
Kamat, et al., 2011[15]	Pathology	Random—8% review	2.6% (35/1,339)
Layfield and Frazier, 2016[14]	Pathology	Random—10% targeted review	0.8% (17/2147)
		Nonrandom—solicited external opinion	7.1% (5/70)
		Nonrandom—unsolicited review by external institution	1.6% (3/190)
		Nonrandom—specific pathology study type	8.5% (5/59)
Raab et al., 2008[12]	Pathology	Random—5% targeted review	2.6% (195/7444)
		Nonrandom—focused review of specific pathology study type	13.2% (50/380)
Swanson et al., 2012[13]	Radiology	Random—mandatory 4 prior comparison studies per shift	3.8% (186/4,892)
		Nonrandom—voluntary review of cases of interest	12% (46/386)

*All cases were selected for review due to the presence of a potential diagnostic error.

Cases selected for review by a random process consistently had lower discrepancy rates between the initial interpretation and the peer review interpretation (0.8%–3.8%) than the cases selected nonrandomly (1.6%–13.2%). The more focused the case selection criteria, the higher the yield of identified diagnostic errors.

In their study of the effectiveness of random and focused reviews in anatomic pathology, Raab et al. found that the 5 percent targeted random selection method identified significantly fewer cases with diagnostic errors than the focused review of case types known to be diagnostically challenging or where there is a lack of standardization (2.6% vs. 13.2%, p<.001).[12] In practical terms, the focused review detected approximately 4 times the number of errors compared with the random reviews, which involved 20 times the number of specimens. Layfield and Frazier compared four different methods of anatomic pathology case selection and found that randomly targeted cases had the lowest rate of identified diagnostic errors (0.8%) compared with the three non-random methods. The focused review, in which all cases of a specific type are reviewed (all dermatopathology cases), identified the greatest percentage of diagnostic errors (8.5%).[14] In a study by Itri et al., at an institution where radiologists are required to review 20 randomly selected cases per month, the discrepancy rate was found to be 2.6 percent, with all identified errors being considered minor discrepancies.[11] The authors also found that, among 190 additional cases selected for review because of concern about potential errors, 130 (68.4%) had significant discrepancies: 94 were significant discrepancies that may affect treatment but not outcomes, and 36 were major discrepancies that may affect outcomes. In a study conducted at a large, urban, multidisciplinary children's hospital, Swanson et al. (2012) describe discrepancy rates of 3.6 percent using their mandatory random peer review process, where each radiologist reviews four cases per shift. Radiologists could also conduct nonrandom reviews on cases of interest or concern, or if they were referenced in a clinical consultation or part of a review conference. There was a 12-percent discrepancy rate using this method.[13]

One study combined the use of random case selection with a prospective review approach, where the peer review occurred prior to report finalization.[15] The rate of discrepancies using this method was 2.6 percent, aligning with the findings of the other studies using random case review. There was one discrepancy considered to be of major significance identified, which was resolved prior to patient care decisions being made.

1.4.3.2 Double Reading

A common form of nonrandom peer review, particularly in radiology practice, is the use of double reading, in which a second clinician reviews a recently completed case.[5] With this method the review is integrated into the diagnostic process rather than conduced retrospectively, allowing errors to be identified and resolved prior to a report being transmitted to the ordering provider or the patient.

Geijer and Geijer (2018) reviewed 46 studies to identify the value of double reading in radiology. The studies fell into two categories: those that used two radiologists of similar degree of subspecialization (e.g., both neuroradiologists) and those that used a subspecialized radiologist only for the second review (e.g., general radiologist followed by hepatobiliary radiologist). Across both types of studies included in the review, double reading increased sensitivity at the expense of reduced specificity. In other words, double reading tended to identify more disease, while also identifying disease in cases that were actually negative (i.e., false positives). With discrepancy rates in studies between 26 and 37 percent, the authors suggest that double reading might be most impactful for trauma CT scans, for which there are a large number of images generated that need to be read quickly under stressful circumstances. The authors also suggest that it may be more efficient to use a single subspecialized radiologist rather than implement double reading, as using a subspecialist as a second reviewer introduced discrepancy rates up to 50 percent. This was thought to be a result of the subspecialist changing the initial reports and the bias introduced by having the subspecialist being the reference standard for the study.[16]

Pow et al. (2016) reviewed 41 studies to assess the use of double reading on diagnostic efficacy for screening and diagnostic imaging studies. As with the previously described systematic review, the use of double reading was found to increase sensitivity and reduce specificity, making it more desirable for tests, such as cancer screening, where high sensitivity is desired.[17] Also consistent with Geijer and Geijer (2018), the authors recommended the use of double reading in trauma due to the large number of images generated and emergent need for results. They also found that the level of expertise of the reviewers influences the error rate, with those review processes using a subspecialist for the second review having higher rates of error detection than those using two radiologists with similar training.

Four studies evaluated the use of double reading, in which the second reading occurred either concurrently with or in immediate proximity to the first reading.[18-21] In each of these studies, significant numbers of discrepancies were determined to be clinically significant by RADPEER scoring criteria. In Agrawal et al, dual reporting identified 145 errors (3.8%; 95% CI, 3.2 to 4.4) that led to report modification, with 69 determined to be clinically significant.[18] Lauritzen et al. identified 146/1,071 (14%, 95% CI, 11.6% to 15.8%) of changes to abdominal CT exam reports that were clinically significant.[19] In a similar study of dual reading for thoracic CT, 91/1,023 (9%) of the report changes were clinically significant, including 3 that were critical and required immediate action and 15 that were major and required a change in treatment.[20] In both studies, the authors found that double reading uncovered errors with less delay and during the time when patient treatment was still able to be affected. Murphy et al. (2010), unlike the other prospective double-reading studies, evaluated blinded double reporting for patients undergoing colon CT scans. They found that, of the 24 significant findings, 7 were identified by only one of the two observers. Although this is counter-intuitive, the probability that a patient with a finding on CT examination had colon cancer was 69 percent for single reporting (11 true positives, 5 false positives) and 54.5 percent for double reporting (12 true positives, 10 false positives). For double reporting, one extra true-positive colon cancer was detected at the expense of five unnecessary colonoscopies (false positives), reducing the positive predictive value.[21]

Lian et al. (2011) compared the diagnostic error rates in a study in which CT angiograms of the head and neck were initially read by a staff neuroradiologist alone (n=144), double-read by staff and a diagnostic radiology resident (n=209), or double-read by staff and a neuroradiology fellow (n=150). Retrospectively, the CT angiograms were then blindly reviewed by two neuroradiologists to detect errors; 503 cases were included, with 26 significant discrepancies discovered in 20/503 studies (4.0%), and all errors were missed diagnoses. Ten of the 26 discrepancies were originally missed by staff alone (6.9% of studies read), 9 by staff and a resident (4.3%), and 7 by staff and a fellow (4.7%). The authors concluded that double reading with a resident or fellow reduces error.[22]

1.4.3.2.1 Economic Outcomes

In their systematic review, Pow et al. (2016) identified six studies from different countries that evaluated cost-effectiveness of double reading for screening mammography. The authors concluded that double reading is a cost-effective strategy. The increased early cancer detection rates outweigh the costs incurred by the double reading, such as infrastructure and additional clinician resources necessary to carry out double reading.[17]

Natarajan et al. (2017) quantified the hospital charges associated with the dual reading by an orthopedist and a radiologist for radiographs at a hospital-based orthopedic clinic. The authors calculated that the total charges for the radiology interpretations for the 2,264 radiographs was $87,362, or $39/study. There were 23 cases where the radiology report provided additional clinically relevant diagnoses not noted by the orthopedist, at the average cost of $3,798 in hospital charges per occurrence.[23]

1.4.3.3 Reinterpretation of Studies Conducted at Outside Institutions

Two studies examined the effect of reinterpretations of radiology studies done at outside institutions. Onwubiko and Mooney (2016) found that out of 98 reinterpreted CT scans of the abdomen and pelvis done in the context of pediatric blunt trauma, 12 significant new injuries were identified, 3 patients had their solid organ injuries upgraded, and 4 patients were downgraded to no injury. The benefit of reinterpreting scans extends beyond identifying potential diagnostic errors to limiting radiation exposure and unnecessary testing in the pediatric population.[24] Lindgren et al. (2014) determined the clinical impact and value of having outside abdominal imaging exams reinterpreted by subspecialized radiologists. Twenty of the 398 report comparison discrepancies (5.0%) had high clinical significance and 30 (7.5%) medium clinical significance. Over half of these discrepancies were due to overcalls, where the outside institution placed more importance on the significance of a finding than was warranted by the second review.[25]

1.4.3.4 Unintended Consequences

1.4.3.4.1 Negative

In the case of dual reading, Natarajan et al. (2017) found that the addition of the radiologist interpretation to the orthopedic interpretation of musculoskeletal films in pediatric orthopedic practice added clinically relevant information in 1.0 percent of the cases, yet misinterpreted 1.7 percent of the cases, potentially adding diagnostic errors into the process.[23] Murphy et al. (2010) found that double reading of colon CT scans increased the number of individuals falsely diagnosed with colon pathology. The protocol found one extra-colonic cancer, but at the expense of five unnecessary endoscopic procedures.[21]

1.4.3.4.2 Positive

Harvey et al. (2016) identified that their group-oriented consensus review method had a secondary effect of fostering a culture of safety in their department, where radiologists feel comfortable identifying and openly discussing diagnostic errors.[10] This finding was supported by Itri et al. (2018), who recognized that peer learning conferences, during which diagnostic errors were reviewed, supported a culture of safety where clinicians learned from their mistakes.[11]

1.4.4 Implementation

1.4.4.1 Facilitators

1.4.4.1.1 Limiting Peer Review to Specific Case Types

Several studies found that certain more complex radiology cases, such as trauma scans or MRIs, benefited more from double reading when compared with examinations such as plain musculoskeletal radiographs.[16,17] Recommendations include the use of subspecialty reinterpretation of high-risk cases, such as in patients with history of cancer or trauma, or using data from peer review to identify areas where there are more likely to be missed diagnoses and focusing peer review on those areas. Raab et al. (2008) recommended a similar approach in pathology, using focused peer review for specific subspecialty cases.[12]

1.4.4.2 Barriers

1.4.4.2.1 Concern Over Maintenance of Confidentiality and Medical Malpractice

Concerns over maintenance of confidentiality by the physicians and fears about the impact of peer review findings on medical malpractice litigation have been identified as a barrier to participation in peer review.[1,26] Several of the studies identified these concerns as barriers to implementing their peer review systems.[10,19] As a way to overcome this challenge, Harvey et al. (2016) described deliberately designing their program to ensure that all information disclosed through the process of peer review is protected under their State's statutory peer review privilege, preventing the information from being used against a clinician in malpractice claims.[10,27] At this time, all 50 States and the District of Columbia have privilege statutes that protect peer review records of medical staff members, although how the privilege is applied may vary by State.

1.4.4.2.2 Increased Staffing Needs, Workload, and Associated Costs

Several studies mentioned the need for increased staffing for peer review activities, requiring additional funds and departmental leadership support and engagement.[10,14-16,21,23,24] One study posited that error rates will depend on the workload of the clinician, with greater workloads leading to greater error rates.[22]

1.4.5 Resources

There are limited resources related to conducting peer review to prevent diagnostic error. As mentioned previously, the ACR offers information regarding RADPEER, their peer review system, on their association's website.[28]

1.4.6 Gaps and Future Directions

Based on the literature identified in this review, traditional random peer review mechanisms employed to maintain compliance with accreditation requirements have not consistently been demonstrated to improve diagnostic accuracy. The studies focus on the rates of discrepancies as detected by the peer review process, but lack follow-through to examine the direct effects on patient harm and clinician performance over time. In addition to uncovering discrepancy rates, there is also a need to identify the root causes of the discrepancies so that they can be understood and prevented. Discrepancies that are generated because of poor image or specimen quality will be addressed very differently from those that are a result of a lack of time or knowledge by the clinician.

The studies did not address the impact of peer feedback, a critical component of peer review.[2,29] There is a missed opportunity to learn from errors, both at the individual and practice levels.[2] It would be beneficial to understand how to best deliver performance feedback and how the feedback is then used to change clinical practice.

There is also a need to design and test different types of peer review systems to maximize their value for improving care while maximizing limited resources. From the literature reviewed, it appears that there is benefit, at least in the field of radiology, to using both random and nonrandom case selection, subspecialist involvement, and prospective and retrospective reviews. Finding the right balance between the different modes of review in terms of clinical effectiveness and cost-effectiveness would be of use.

The available literature on peer review and its impact on diagnosis is focused on the fields of pathology and radiology, areas where peer review has been used the longest as part of quality assurance programs. It would be valuable to expand the breadth of studies to include other forms of peer review, including group consensus or conferences that could potentially be used to improve diagnostic accuracy in other fields, such as primary care, as these methods might be suitable for diagnostic dilemmas encountered in a variety of settings.

Lastly, even with the use of peer review in any of its forms, patients continue to experience errors in diagnostics, some significant and with a real potential for harm. In the Improving Diagnosis in Health Care report, the Committee on Diagnostic Error in Health Care, after reviewing the evidence, concluded that "most people will experience at least one diagnostic error in their lifetime, sometimes with devastating consequences."[3] This is disconcerting and speaks to the need for considering "upstream" measures as well—not just relying on the inspection mode at the point of care but also looking at re-engineering the entire process for more- accurate diagnosis.[30] In order to start this process, efforts should be directed to elucidate the root causes of diagnostic errors. This knowledge can then be used to guide the development of strategies aimed at improving the underlying system of care.

References for Section 1.4

1. Kaewlai R, Abujudeh H. Peer review in clinical radiology practice. Am J Roentgenol. 2012;199(2):W158-W62.
2. Butler GJ, Forghani R. The next level of radiology peer review: enterprise-wide education and improvement. J Am Coll Radiol. 2013;10(5):349-53.
3. National Academies of Sciences, Engineering and Medicine. 2015. Improving Diagnosis in Health Care. Washington, DC: National Academies Press. https://doi.org/10.17226/21794.
4. Commission TJ. Ongoing Professional Practice Evaluation (OPPE) - Low Volume Practitioners - Data Use From Another Organization. https://www.jointcommission.org/en/standards/standard-faqs/critical-access-hospital/medical-staff-ms/ka02s000000loqc/.
5. Moriarity AK, Hawkins CM, Geis JR, et al. Meaningful peer review in radiology: a review of current practices and potential future directions. J Am Coll Radiol. 2016;13(12):1519-24.
6. Larson DB, Donnelly LF, Podberesky DJ, et al. Peer feedback, learning, and improvement: answering the call of the Institute of Medicine report on diagnostic error. Radiol. 2016;283(1):231-41.
7. Schiff GD, Ruan EL. The Elusive and lllusive quest for diagnostic safety Metrics. J Gen Intern Med 2018;33(7):983-5.10.1007/s11606-018-4454-2. https://doi.org/10.1007/s11606-018-4454-2.
8. Mucci B, Murray H, Downie A, et al. Interrater variation in scoring radiological discrepancies. The Brit J Radiol. 2013;86(1028):20130245.
9. Goldberg-Stein S, Frigini LA, Long S, et al. ACR RADPEER committee white paper with 2016 updates: revised scoring system, new classifications, self-review, and subspecialized reports. J Am Coll Radiol. 2017;14(8):1080-6.
10. Harvey HB, Alkasab TK, Prabhakar AM, et al. Radiologist peer review by group consensus. J Am Coll Radiol. 2016;13(6):656-62.
11. Itri JN, Donithan A, Patel SH. Random versus nonrandom peer review: a case for more meaningful peer review. J Am Coll Radiol. 2018;15(7):1045-52.
12. Raab SS, Grzybicki DM, Mahood LK, et al. Effectiveness of random and focused review in detecting surgical pathology error. Am J Clin Pathol. 2008;130(6):905-12.
13. Swanson JO, Thapa MM, Iyer RS, et al. Optimizing peer review: a year of experience after instituting a real-time comment-enhanced program at a children’s hospital. Am J Roentgenol. 2012;198(5):1121-5.
14. Layfield LJ, Frazier SR. Quality assurance of anatomic pathology diagnoses: Comparison of alternate approaches. Pathol Res Pract. 2017;213(2):126-9.
15. Kamat S, Parwani AV, Khalbuss WE, et al. Use of a laboratory information system driven tool for pre-signout quality assurance of random cytopathology reports. J Pathol Inform. 2011;2.
16. Geijer H, Geijer M. Added value of double reading in diagnostic radiology, a systematic review. Insights Imaging. 2018;9(3):287-301.
17. Pow RE, Mello-Thoms C, Brennan P. Evaluation of the effect of double reporting on test accuracy in screening and diagnostic imaging studies: a review of the evidence. J Med Imag Radiat On. 2016;60(3):306-14.
18. Agrawal A, Koundinya D, Raju JS, et al. Utility of contemporaneous dual read in the setting of emergency teleradiology reporting. Emerg Radiol. 2017;24(2):157-64.
19. Lauritzen PM, Andersen JG, Stokke MV, et al. Radiologist-initiated double reading of abdominal CT: retrospective analysis of the clinical importance of changes to radiology reports. BMJ Qual Saf. 2016;25(8):595-603.

20. Lauritzen PM, Stavem K, Andersen JG, et al. Double reading of current chest CT examinations: Clinical importance of changes to radiology reports. Euro J Radiol. 2016;85(1):199-204.
21. Murphy R, Slater A, Uberoi R, et al. Reduction of perception error by double reporting of minimal preparation CT colon. Br Journal Radiol. 2010;83(988):331-5.
22. Lian K, Bharatha A, Aviv R, et al. Interpretation errors in CT angiography of the head and neck and the benefit of double reading. Am J Neuroradiol. 2011;32(11):2132-5.
23. Natarajan V, Bosch P, Dede O, et al. Is there value in having radiology provide a second reading in pediatric orthopaedic clinic? J Pediatr Orthop. 2017;37(4):e292-e5.
24. Onwubiko C, Mooney DP. The value of official reinterpretation of trauma computed tomography scans from referring hospitals. J Pediatric Surg. 2016;51(3):486-9.
25. Lindgren EA, Patel MD, Wu Q, et al. The clinical impact of subspecialized radiologist reinterpretation of abdominal imaging studies, with analysis of the types and relative frequency of interpretation discrepancies. Abdominal Imaging. 2014;39(5):1119-26.
26. Mahgerefteh S, Kruskal JB, Yam CS, et al. Peer review in diagnostic radiology: current state and a vision for the future. Radiographics. 2009;29(5):1221-31.
27. Commonwealth of Massachusetts General Laws. https://malegislature.gov/laws/generallaws
28. Radiology ACo. RADPEER. https://www.acr.org/Clinical-Resources/RADPEER.
29. Sheu YR, Feder E, Balsim I, et al. Optimizing radiology peer review: a mathematical model for selecting future cases based on prior errors. J Am Coll Radiol. 2010;7(6):431-8.
30. Berwick DM. Continuous Improvement as an Ideal in Health Care. New Engl J Med. 1989;320(1):53-6.10.1056/nejm198901053200110. https://www.nejm.org/doi/full/10.1056/NEJM198901053200110.

Conclusion and Comment

The PSPs reviewed in this chapter aim to reduce diagnostic errors by targeting cognitive-based factors and systems-based factors. The evidence in support of these practices varied in depth and consistency.

CDS offers solutions to address diagnostic errors through incorporation of evidence-based diagnostic protocols, and improve communication and integration with clinical workflow. This review found that CDS may improve diagnosis, although the studies tend to be exploratory in nature, validating the decision algorithms. The use of AI and machine learning has generated excitement over its potential, but they are also exploratory and lack testing during the care of actual patients. These systems need to be reassessed once fully implemented and iteratively improved in real clinical settings on patients actively undergoing diagnosis. Studies included in the review also support the notion that CDS tools are best used as adjuncts to the clinician's decision making process and not as replacements. This was particularly true for CDS tools that assist with diagnostic study interpretation, such as ECG interpretation. The literature also identified that the diagnoses generated by CDS tools are only as good as the information that is put into the system; if the initial assessment of the patient (e.g., physical exam finding) is incorrect, it is likely that the output will be incorrect.

RNSs aim to address lapses in communication, a contributing factor to delayed diagnosis and treatment of patients in both ambulatory and inpatient settings. There was considerable variability in the findings of the included studies, with the results being dependent on many factors, including the type of the test, the type of communication (i.e., synchronous or asynchronous), and whether the alert was manual or automated. Studies were conducted in a surprisingly limited number of institutions. For both critical and non-critical CSTR of radiologic studies, lab studies, and tests pending at discharge, the use of RNS showed mixed results in the timeliness of receipt and in action on the test results. Policies and procedures that aligned with the system, mindful integration of the RNS into the existing workflow, and appropriate staffing were identified as factors supporting successful implementation of the systems. Barriers to successful implementation, particularly when results are conveyed across transitions from inpatient to outpatient settings, include the lack of connectivity between hospitals and non-network physicians. Additionally, there were operational challenges associated with providing critical alerts to physicians who may not be available at the time the result is available (e.g., not on call). Ultimately, they have a central role to play in closed-loop systems to ensure reliability and tracking of critical test results.

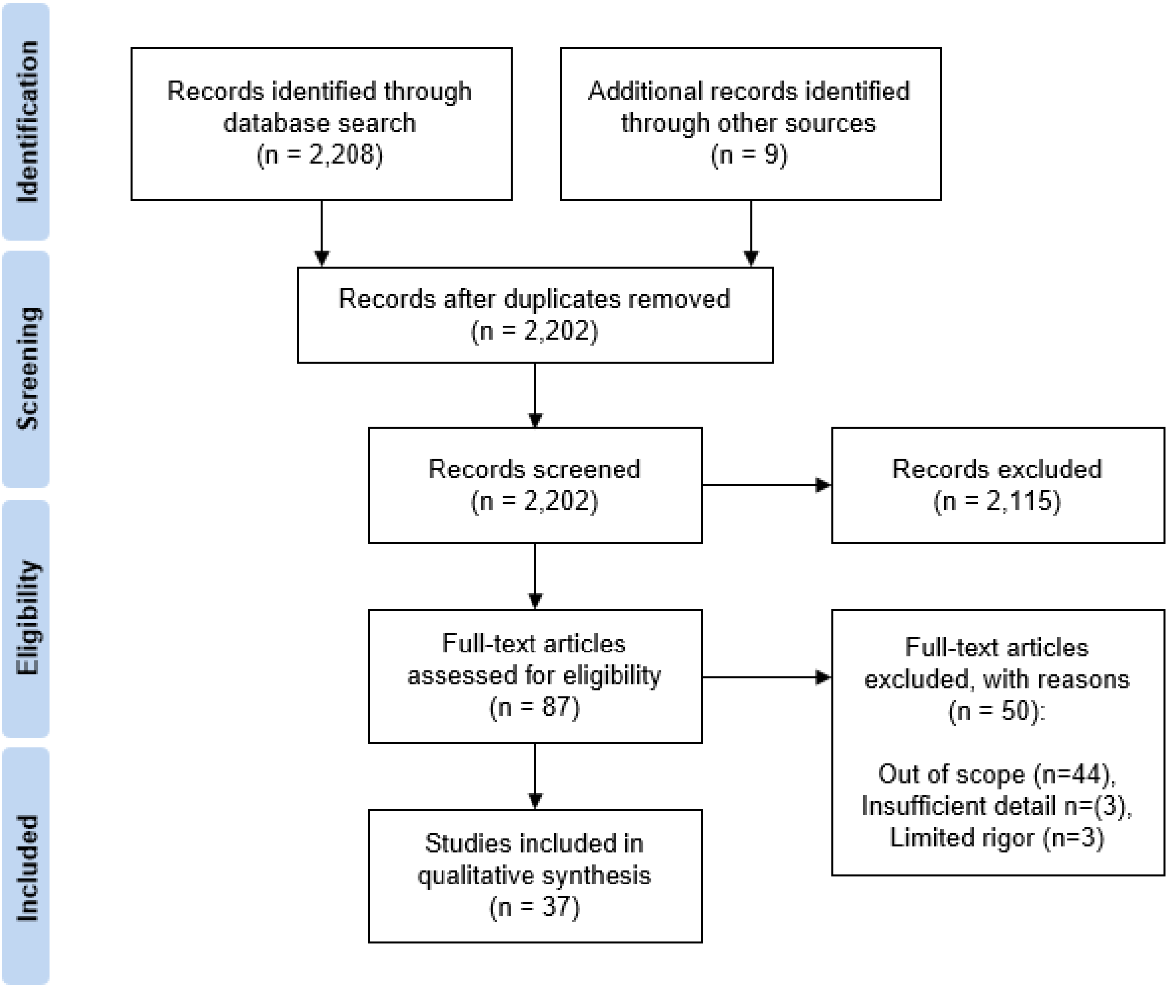
Evidence to support education and training on the diagnostic process to enhance clinical reasoning and decrease biases showed generally positive results, with study designs being strong (e.g., randomized controlled trials), although there was some lack of generalizability, as many of the studies had low numbers of subjects. Training on metacognitive skills as a way to reduce biases may improve diagnostic accuracy, particularly as clinical experience increases. Online training, either didactic or simulation based, was shown to be successful at improving clinical reasoning skills. Of note, there was a dearth of studies that examined the transfer of learning from classroom or simulated settings into the clinical setting and actual patient care, where there is a critical need for future research.

For the PSP of peer review, studies show significant numbers of missed or misread test interpretations. However, there is a lack of evidence to show that traditional random peer review and feedback mechanisms used in radiology or pathology to maintain compliance with accreditation requirements improve diagnostic quality over time or prevent diagnostic errors from reaching the patient. For both radiology and pathology, nonrandom peer review appears to be more effective at identifying diagnostic errors than random peer review; and when peer review is conducted prospectively, there is an opportunity to identify diagnostic errors before they reach or harm the patient.

Since the previous Making Health Care Safer Report was published, studies examining the use of these four PSPs are increasing in both number and quality. Overall, there is still a relative dearth of studies focused on diagnostic error prevention and methods to improve diagnostic accuracy compared with other patient safety topics. General considerations for future research in diagnostic safety include the use of consistent measures and definitions of diagnostic error to allow comparisons of studies and aggregation of data across smaller studies (i.e., meta-analyses), moving from exploratory studies to studies conducted in real clinical settings in real time, and understanding how to best integrate technology with the current workflow to support diagnosis-related activities. There is also a need to design and test innovative and more refined versions of the past interventions using more advanced educational, quality improvement, and health information technology tools in the future.

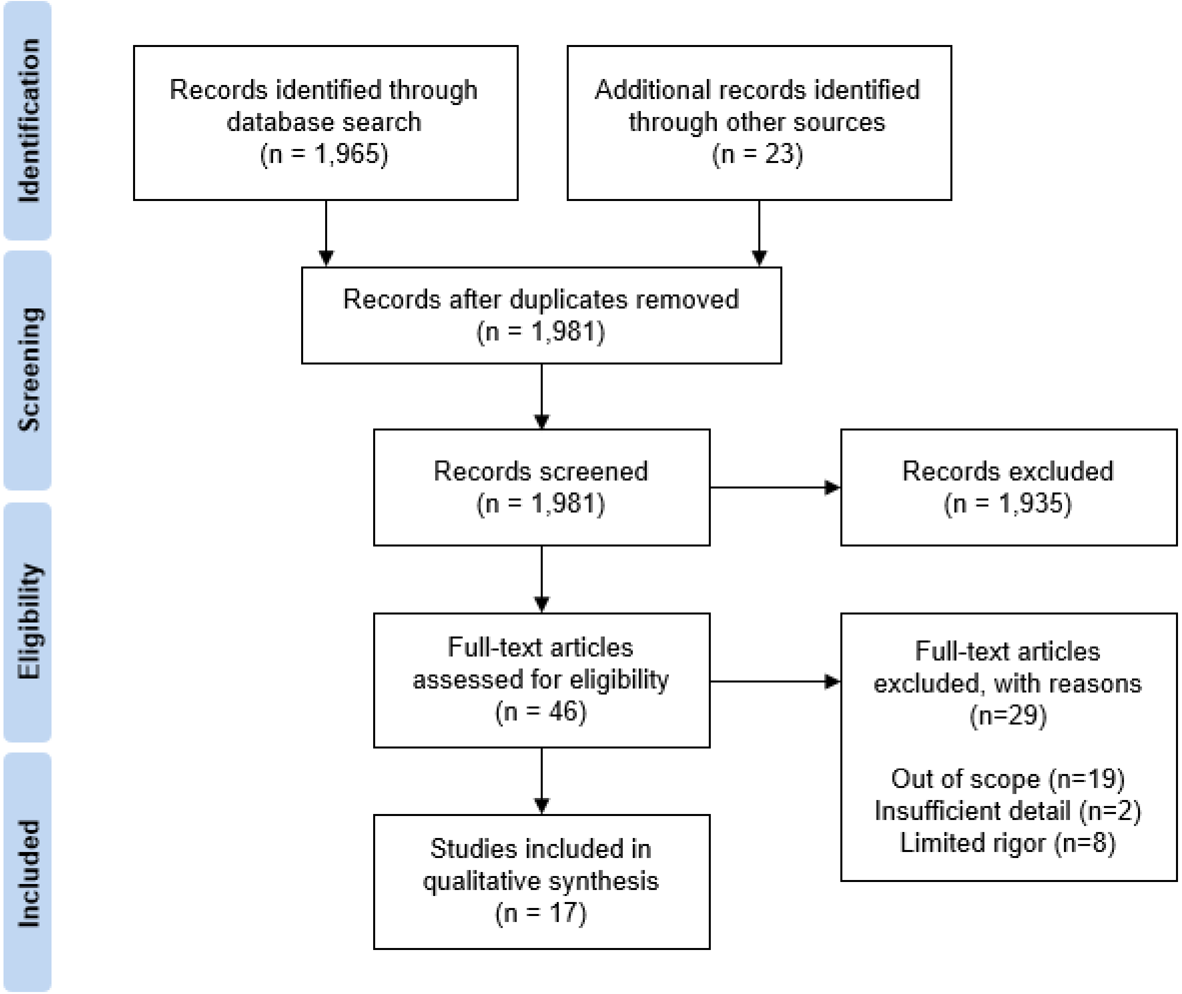
Appendix A. Diagnostic Errors PRISMA Diagrams

Figure A.1: Diagnostic Errors, Clinical Decision Support—Study Selection for Review


Identification
Screening
Eligibility
Included
Records identified through database search (n = 2,208)
Additional records identified through other sources (n = 9)
Records after duplicates removed (n = 2,202)
Records screened (n = 2,202)
Records excluded (n = 2,115)
Full-text articles assessed for eligibility (n = 87)
Full-text articles excluded, with reasons (n = 50):
Out of scope (n=44), Insufficient detail n=(3), Limited rigor (n=3)
Studies included in qualitative synthesis (n = 37)

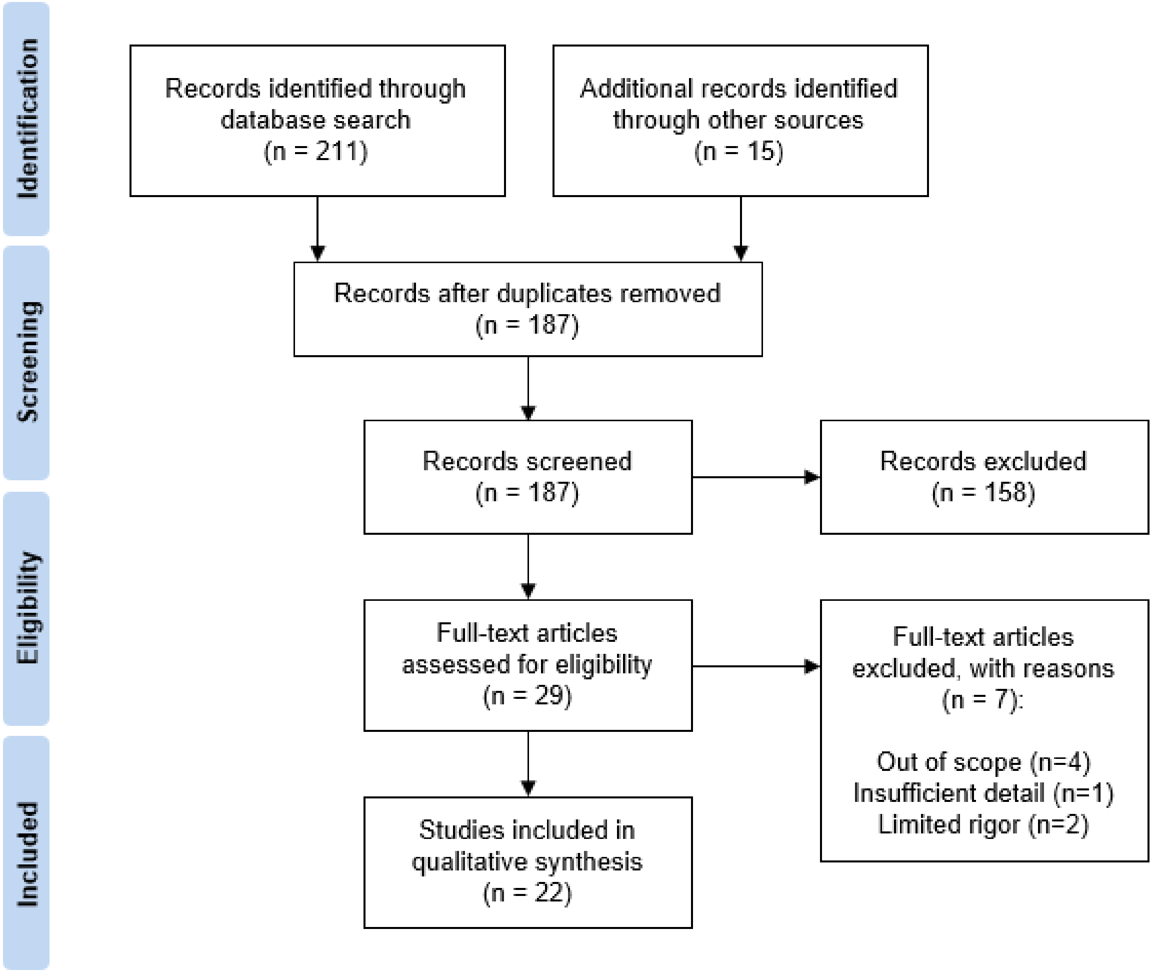
PRISMA criteria described in Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7): e1000097. doi:10.1371/journal.pmed1000097.

Figure A.2: Diagnostic Errors, Result Notification Systems—Study Selection for Review


Identification
Screening
Eligibility
Included
Records identified through database search (n = 1,965)
Additional records identified through other sources (n = 23)
Records after duplicates removed (n = 1,981)
Records screened (n = 1,981)
Records excluded (n = 1,935)
Full-text articles assessed for eligibility (n = 46)
Full-text articles excluded, with reasons (n=29)
Out of scope (n=19)
Insufficient detail (n=2)
Limited rigor (n=8)
Studies included in qualitative synthesis (n = 17)

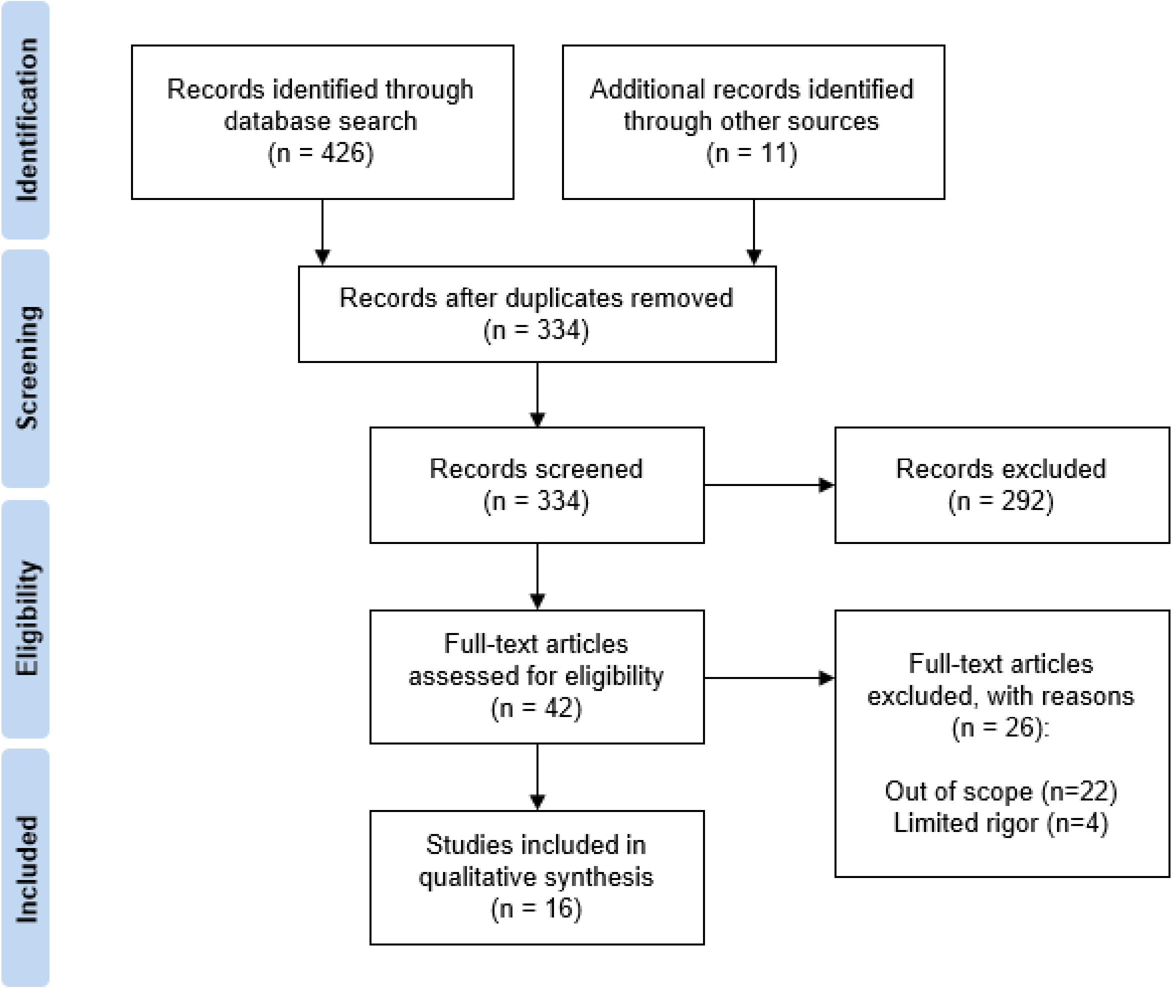
PRISMA criteria described in Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7): e1000097. doi:10.1371/journal.pmed1000097.

Figure A.3: Diagnostic Errors, Education and Training—Study Selection for Review


Identification
Screening
Eligibility
Included
Records identified through database search (n = 211)
Additional records identified through other sources (n = 15)
Records after duplicates removed (n = 187)
Records screened (n = 187)
Records excluded (n = 158)
Full-text articles assessed for eligibility (n = 29)
Full-text articles excluded, with reasons (n = 7):
Out of scope (n=4)
Insufficient detail (n=1)
Limited rigor (n=2)
Studies included in qualitative synthesis (n = 22)

PRISMA criteria described in Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7): e1000097. doi:10.1371/journal.pmed1000097.

Figure A.4: Diagnostic Errors, Peer Review—Study Selection for Review


Identification
Screening
Eligibility
Included
Records identified through database search (n = 426)
Additional records identified through other sources (n = 11)
Records after duplicates removed (n = 334)
Records screened (n = 334)
Records excluded (n = 292)
Full-text articles assessed for eligibility (n = 42)
Full-text articles excluded, with reasons (n = 26):
Out of scope (n=22)
Limited rigor (n=4)
Studies included in qualitative synthesis (n = 16)

PRISMA criteria described in Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7): e1000097. doi:10.1371/journal.pmed1000097.

Appendix B. Diagnostic Errors Evidence Tables

Table B.1: Diagnostic Errors, Clinical Decision Support—Single Studies

Note: Full references are available in the Section 1.1 reference list.

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Arthi et al., 2008[22]	A neuro-fuzzy system, using both artificial neural network (ANN) and fuzzy logic models, designed for the identification or diagnosis of autism	Evaluation of model performance; 194 samples.	Not specified	In this neuro-fuzzy model, the network was shown to learn quickly, and has an output error rate of 0.01, which remained constant after 400 epochs. The overall performance of this model is 85–90%, aiding in the diagnosis of autism.	Not provided	Low	None
Bien et al., 2018[27]	Automated deep learning model for detecting general abnormalities and specific diagnoses on knee magnetic resonance imaging (MRI) scans	Evaluation of model performance; internal validation using 1,370 knee MRIs performed between January 1, 2001, and December 21, 2002 (Stanford Univ. Medical Center). External validation using public dataset of 917 knee MRI exams (Clinical Hospital Center, Rijeka, Croatia).	Stanford University Medical Center, United States; Clinical Hospital Centre, Rijeka, Croatia	The model achieved area under the receiver operating characteristic curve (AUC) values of 0.937 (95% confidence interval [CI], 0.895 to 0.980) in detecting general abnormalities, 0.965 (95% CI, 0.938 to 0.993) for ACL tears, and 0.847 (95% CI, 0.780 to 0.914) for meniscal tears. Authors found no significant differences between the performance of the model and that of unassisted general radiologists in detecting abnormalities. Providing model predictions significantly increased clinical experts' specificity in identifying ACL tears ($p<0.001$; q-value 0.006).	Not provided	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Bond et al., 2012[5]	Differential diagnosis (DDX) generator	Analysis of performance of four DDX programs (Diagnosis Pro, DXPlain, Isabel, PEPID) using 20 test cases.	Not specified	The mean scores (95% CI) from performance testing on a five-point scale were Isabel 3.45 (2.53 to 4.37), DxPlain 3.45 (2.63 to 4.27), Diagnosis Pro 2.65 (1.75 to 3.55), and PEPID 1.70 (0.71 to 2.69).	Integration with electronic health record (EHR)—at the time of the publication, the DDX were limited by the data fields shared with the EHR. Better integration of the systems with the EHR would overcome this challenge.	Moderate	Included in Riches 2016, systematic review and meta-analysis
Cairns et al., 2017[32]	Electrocardiogram (ECG) interpretation support system (interactive progressive-based interpretation [IPI] system and differential diagnosis algorithm [DDA]) designed to augment the human interpretation process	Counterbalanced trial using convenience sampling; 35 participants completing 375 interpretations (215 control, 160 using support); training levels of subjects include medical students through cardiologists.	Classroom environment and remotely via website hyperlinks	IPI + DDA approach was shown to improve diagnostic accuracy by 8.7% (although this was not statistically significant). The percentage of correct interpretations for reading ECGs using the conventional approach was 42.61%. Interpretations using the IPI + DDA method were 51.35% (chi-squared p-value=0.1852).	Not provided	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Chamberlain et al., 2016[19]	Mobile smart phone application that consists of an electronic stethoscope, a peak flow meter application, and a patient questionnaire. Data from the app are combined with a machine-learning algorithm to identify patients with asthma and chronic obstructive pulmonary disease (COPD)	Evaluation of model performance; 119 healthy and sick participants used the app and also were examined by an experienced pulmonologist using a full pulmonary testing laboratory.	Not specified	Employing a two-stage logistic regression model, the algorithms were first able to identify patients with either asthma or COPD from the general population, yielding an AUC of 0.95. Then, the algorithm was able to distinguish between patients with asthma and patients with COPD, yielding an AUC of 0.97.	Not provided	Moderate	None
Chou et al., 2017[15]	Visually based, computerized diagnostic decision support system (VCDDSS, VisualDx)	Pre/post study design, no comparison group. Clinical diagnoses of 13 patients were made by 51 sixth-year medical students, 13 dermatology residents, and one consultant dermatologist.	Dermatology Teaching Clinic, China Medical University Hospital, Taiwan	There was an 18.75% increase in diagnostic accuracy after use of VCDDSS (accuracy rate before using VCDDSS 62.5%, after VCDDSS 81.25%; p<0.01).	Not provided	Moderate	None
David et al., 2011[9]	Visually based, computerized diagnostic decision support system (VCDDSS, VisualDx)	Descriptive analysis of model performance; 80 patients admitted with a diagnosis of cellulitis.	Harbor-UCLA Medical Center, United States	Twenty-eight out of 80 cases admitted for cellulitis had alternative diagnoses (i.e., were misdiagnoses). The admitting physician included the correct diagnosis in the DDX in 4/28 (14%) and the VCDDSS in 18/28 (64%) of the misdiagnosed cases (p= 0.0003).	Not provided	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Deleger et al., 2013[17]	Natural language processing (NLP) and machine-learning (ML) based automated method to risk stratify abdominal pain patients by analyzing the content of the EHR	Retrospective observational study; 2,100 pediatric emergency department patients with abdominal pain.	Pediatric emergency department (ED) in an urban, quaternary care children's hospital, United States	The system performance was comparable to that of physician experts, and achieved an average F- measure of 0.867 (recall or sensitivity, 0.869; precision or PPV, 0.863) for risk classification.	Not provided	Low-moderate	None
Elkin et al., 2010[39]	DXplain, a computer-based medical education, reference, and decision support system	Pre/post study design; residents doing month-long rotations on one of five general medicine services; 323 uses of the DXplain in the post-intervention period.	General medicine services at St. Mary's Hospital, a 1,200-bed hospital operated by the Mayo Clinic, Rochester, MN, United States	Five hundred sixty-four cases were identified as diagnostically challenging by the criteria during the intervention period, along with 1,173 cases during the control period. Total charges were $1,281 lower (p=.006), Medicare Part A charges $1,032 lower (p=.006), and cost of service $990 lower (p=.001) per admission in the intervention cases than in control cases.	Not provided	Low-moderate	Included in Riches, 2016, systematic review and meta-analysis
Farmer, 2014[25]	Diagnostic clinical decision support system (CDS) developed to assist primary care clinicians in diagnosing musculoskeletal shoulder complaints and to reduce diagnostic errors	Prospective observational audit; 93 patients attending the Shoulder Clinic between June and December 2012.	Orthopedic outpatient department at the Royal Hampshire County Hospital, part of the Hampshire Hospitals NHS Foundation Trust, United Kingdom	CDS showed significant high levels of sensitivity (91%), specificity (98%), positive likelihood ratio (53.12), and negative likelihood ratio (0.08), with a kappa value of 0.88 to a confidence level of 99% compared with expert diagnosis combined with arthroscopy findings or radiological imaging.	Not provided	Low-moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Gegundez-Fernandez et al., 2017[10]	A mobile app-based decisions support system for the differential diagnosis of uveitis	Retrospective case-series study; a series of 159 patients originally diagnosed by a uveitis specialist with specific uveitis (N=88) and idiopathic uveitis (N=71).	Two hospitals in Madrid, Spain	Diagnostic accuracy of the CDS was 96.6% (95% CI, 93.2 to 100).	The successful use of DDSS is fully dependent on proper assessment of symptoms and signs by the responsible clinician, because the computer will process only the data the human introduces.	Moderate	None
Graber et al., 2008[3]	Web-based CDS that accepts either key findings or whole-text entry and uses a novel search strategy to identify candidate diagnoses from the clinical findings	Descriptive analysis of model performance; tested 50 consecutive internal medicine adult medical case studies published in the New England Journal of Medicine.	Not specified	The clinical decision support system suggested the correct diagnosis in 48 of 50 cases (96%) with key findings entry, and in 37 of the 50 cases (74%) if the entire case history was pasted into the system.	Not provided	Moderate	Included in Riches. 2016 systematic review and meta-analysis
Gulshan et al., 2016[28]	Deep learning-trained algorithm for automated detection of referable diabetic retinopathy (RDR) and diabetic macular edema in retinal fundus photographs	Algorithm trained using a retrospective development data set of 128,175 retinal images, and validated using 2 separate datasets, both graded by at least 7 U.S. board-certified ophthalmologists.	Not specified	For RDR, the algorithm had an area under the receiver operating curve of 0.991 (95% CI, 0.988-0.993) for the first validation dataset and 0.990 (95% CI, 0.986-0.995) for the second validation dataset.	Not provided	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Hakacova et al., 2012[33]	Computer-based rhythm analysis software—Philips Medical (Software A) and Draeger Medical (Software B)	Descriptive analysis of model performance; 500 ECGs were analyzed manually by two senior experts and three non-expert clinicians, and automatically by two automated systems.	Emergency department, Lund University Hospital, Sweden	Accuracy of nonexpert reading was 85%, not significantly different when compared with the accuracies of the system readings of 80% for system A (p= 45) and 75% for system B (p=.11).	Not provided	Moderate	None
Herweh et al., 2016[29]	e-ASPECTS, a machine learning algorithm that is based on the Alberta Stroke Program Early CT score (ASPECTS), an established 10-point quantitative topographic computed tomography scan score to detect stroke on CT scans	Evaluation of model performance; images of 34 patients with stroke between January 2005 and December 2015; studies interpreted by three stroke experts and three neurology residents.	University Hospital, Heidelberg, Germany	e-ASPECTS showed a similar performance to that of stroke experts in the assessment of brain computed tomography (CT) scans of acute ischemic stroke patients with the Alberta Stroke Program Early CT score method.	Not provided	Moderate	None
Hughes et al., 2017[31]	Automated ECG computerized analysis	Prospective cohort study; 855 triage ECGs obtained between November 14, 2014, and March 3, 2015.	Adult ED, University of North Carolina, United States	A total of 222 (26%) ECGs were interpreted by the computer as normal. The negative predictive value for triage ECGs interpreted by the computer as "normal" was calculated to be 99% (95% confidence interval= 97 to 99).	Not provided	Low-moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Kharbanda et al., 2016[18]	Electronic CDS tool that includes three components: a standardized abdominal pain order set, a web-based risk stratification tool, and a "time of ordering alert"	Quasi-experimental study; 2,803 children age 3 to 18 years who presented with possible appendicitis to the pediatric emergency department (ED) between January 2011 and December 2013.	Two urban, tertiary care pediatric EDs, United States	Use of the CDS tool led to a 54% relative decrease in CT use, with an increase in ultrasound use. No differences in rates of missed appendicitis, ED revisits within 30 days, appendiceal perforation, or ED length of stay between time periods	Not provided	Low	None
Koopman et al., 2015[37]	Machine-learning algorithm-based system designed to match final radiology reports to final ED diagnosis to identify potentially missed diagnoses of fractures.	Evaluation of model performance; 2,378 free-text radiology reports of limb structures.	EDs of three large Australian public hospitals (adults, children, and mixed adults/children)	The PPV (precision) for all data sets=.92; sensitivity (recall)=.92, F-measure=0.92.	The reconciliation process is affected by the way ICD-10 codes are assigned, with many flagged cases being situations in which the abnormality was known but was not conveyed in the assigned ICD-10 code.	Low-moderate	None
Kostopoulou et al., 2017[14]	Prototype CDS integrated in an EHR system and designed to support a clinician's initial assessment by generating a list of possible diagnoses as the reason for encounter (RfE) is entered into the system	Within-subject study design using 12 manufactured scenarios with standardized patients, four for each of the available RfE.	Kings College, London, United Kingdom	Improvement in diagnosis using the CDS was statistically significant (odds ratio [OR] 1.41; 95% CI, 1.13 to 1.77; p=0.003), as were the improvements in diagnostic certainty and management.	Not provided	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Lee et al., 2013[23]	Preclustering-based ensemble learning (PEL) technique to assist in the diagnosis of acute appendicitis	Evaluation of model performance; 574 appendectomy cases, of which 110 were negative for appendicitis.	Tertiary hospital in southern Taiwan	The PEL technique had the best overall performance of classification systems and scoring systems, with an area under the curve measure of 0.619. PEL is more sensitive to identifying positive acute appendicitis than the commonly used Alvarado scoring system, and exhibits higher specificity in identifying negative acute appendicitis.	Not provided	Low	None
Li et al., 2018[30]	Endoscopic images-based nasopharyngeal malignancy detection model (eNPM-DM)	Evaluation of model performance; 27,536 biopsy-proven images from 7,951 individuals obtained from January 1, 2008, to December 31, 2016, split into the training, validation, and test sets; 1,430 images obtained from January 1, 2017, to March 31, 2017, used as a prospective test set.	Sun Yat-sen University Cancer Center; Guangzhou, China	The eNPM-DM attained an overall accuracy of 88.7% (95% CI, 87.8 to 89.5) in detecting malignancies in the test set. In the prospective comparison phase, eNPM-DM outperformed the experts: the overall accuracy was 88.0% (95% CI, 86.1 to 89.6%) versus 80.5% (95% CI, 77.0 to 84.0).	Not provided	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Lin et al., 2009[24]	Intelligent diagnosis model using classification and regression tree (CART) and case-based reasoning (CBR) techniques to increase the accuracy of liver disease diagnosis	Evaluation of model performance; 510 outpatients (300 with liver disease; 210 without) from 2005 to 2006.	Medical Center, Taiwan	Comparing the receiver operating characteristic (ROC) curves of these two models, CART demonstrated a greater sensitivity (0.931) for any given specificity than CBR (0.857). These results suggest the use of CART over CBR for the classification of liver disease. Tested by accuracy, sensitivity, and specificity, CART reports a greater classification capability than does CBR.	Not provided	Low-moderate	None
Martinez-Franco et al., 2018[13]	DXplain, a computer-based medical education, reference, and decision support system	Randomized controlled trial; 87 first-year family medicine residents (44 control, 43 intervention), solving 30 clinical diagnosis cases.	National Autonomous University of Mexico (UNAM) Postgraduate Studies Division in Mexico City, Mexico	There was a significant difference between the percent-correct scores for the control group (74.1±9.4) and the DXplain intervention group (82.4±8.5, p<0.001).	Not provided	Low-moderate	None
Mawri et al., 2016[34]	Computer-interpreted ECG (cECG)	Retrospective cohort study; 340 consecutive patients from September 2003 to December 2009 with STEMI who underwent emergent cardiac catheterization and percutaneous coronary intervention.	Henry Ford Hospital. Detroit, MI, United States	cECG failed to identify 30% of patients with STEMI. Protocol using the immediate review of ECGs by an emergency physician rather than depending on the cECG interpretation led to faster activation of the catheterization laboratory {19 minutes [interquartile range (IQR): 10–37] versus 16 minutes [IQR: 8–29]; p<0.029} and in median door-to-balloon times {113 minutes [IQR: 86–143] versus 85 minutes [IQR: 62–106]; p<0.001} in patients with STEMI.	If there are issues with the recording (e.g., incorrect lead placement, movement artifacts), the accuracy of the cECG interpretation will be affected.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Murphy et al., 2015[38]	Electronic triggers to identify patients at risk of diagnostic delays based on the following criteria: presence of a clinical clue or red flag; exclusion of records where further evaluation is not warranted (e.g., terminal illness); and presence of delay in diagnostic evaluation	Cluster randomized controlled trial; 72 full-time primary care providers (36 in control group, 36 in intervention group) seeing an estimated 118,400 patients in internal or family medicine ambulatory clinics from April 20, 2011, to July 19, 2012.	Urban Veterans Affairs facility (site A) and a private health system (site B), United States	Of 10,673 patients with abnormal findings, the trigger flagged 1,256 patients (11.8%) as high risk for delayed diagnostic evaluation. Times to diagnostic evaluation were significantly lower in intervention patients compared with control patients flagged by the colorectal trigger (median, 104 vs. 200 days, n= 557; p<.001) and prostate trigger (40% received evaluation at 144 vs. 192 days, n=157; p<.001) but not the lung trigger (median, 65 vs. 93 days, n=19; p=.59). More intervention patients than control patients received diagnostic evaluation by final review (73.4% vs. 52.2%, relative risk, 1.41; 95% CI, 1.25 to 1.58).	Not provided	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Niemi et al., 2009[16]	CDS, the Core Measure Manager (CCM), to identify core measure patients (HF and pneumonia) in real time and to provide alerts to the appropriate clinician with sufficient time to allow for intervention when performance measures were not being met	Descriptive analysis of system performance. Pneumonia study: patients 18 years and older with an ED visit, hospital admission, or both between October 1, 2006, and October 31, 2006 (986 admissions, 37 with pneumonia); heart failure (HF) study: patients 18 years and older admitted between February 11, 2007, and March 12, 2007 (1,037 admissions, 94 with HF).	Sutter Medical Center, Sacramento, CA, United States	The sensitivity for identification of pneumonia using the CDS in the ED was 89% and the specificity was 86%. The sensitivity for pneumonia admissions was 92% and the specificity was 90%. The sensitivity for HF identification was 94% and the specificity was 90%.	Not provided	Moderate	None
Segal et al., 2014[11]	CDS, SimulConsult, which generates different diagnoses based on input patient clinical findings	Evaluation of CDS using pre/post design; 16 pediatric neurologists (11 in the final year of pediatric neurology residency or subsequent year [‘‘junior’]) and 5 in practice for >10 years [‘‘senior’’]) tested 40 written case vignettes of patients with neurogenetic diagnoses.	Not specified	Diagnostic errors after using the decision support (‘‘aided’’) fell from 36% to 15% overall. There was an increase in the relevance of listed differential diagnoses after using the software (p< .0001).	A key factor that improved performance was taking enough time (>2 minutes) to enter clinical findings into the software accurately.	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Segal et al., 2016[12]	CDS, SimulConsult, which generates different diagnoses based on input patient clinical findings	Evaluation of CDS using pre/post design. Twenty-six testers (7 general pediatrics, 9 emergency medicine, 10 pediatric rheumatology), eight case vignettes of real patients with confirmed diagnoses (six had pediatric rheumatologic diagnoses; two had other conditions with some rheumatologic findings).	Not specified	Significant reduction in diagnostic errors following introduction of the CDS, from 28% errors to 15% using decision support (p< 0.0001). Improvement was greatest for emergency medicine physicians (p= 0.013) and clinicians in practice for less than 10 years (p= 0.012).	Testers spent an average of 20 minutes per case, of which half was spent using the decision support.	Moderate	None
Song et al., 2016[26]	CDS, based on an online algorithm, that incorporates contextual information and makes diagnostic recommendations to physicians, aiming to minimize the false positive rate of breast cancer diagnosis, given a predefined false negative rate	Evaluation of the CDS algorithm using a de-identified dataset of 4,640 individuals who underwent screening and diagnostic mammograms at a large academic medical center.	Large academic medical center	Proposed approach outperforms the current clinical practice by 36% in terms of false positive rate given a 2% false negative rate.	Not provided	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Vandenberghe et al., 2017[35]	Computer-aided diagnosis using a convolutional neural network model (ConvNets) that automatically scores HER2, a biomarker that defines patient eligibility for anti-HER2 targeted therapies in breast cancer	Evaluation of model performance using a cohort of 71 breast tumor resection samples.	Not specified	In a cohort of 71 breast tumor resection samples, automated scoring showed a concordance of 83% with a pathologist. The 12 discordant cases were then independently reviewed, leading to a modification of diagnosis from initial pathologist assessment for 8 cases.	Not provided	Low-moderate	None
Wolf et al., 2013[20]	Four smartphone applications that allow the use of existing images of skin lesions to make assessments on the likelihood of malignancy risk	Case-control diagnostic accuracy study; a total of 188 lesions evaluated using the four applications (60 melanomas: 44 invasive and 16 in situ; 128 benign lesions).	Not specified	Sensitivity of the four tested applications ranged from 6.8% to 98.1%. Specificity ranged from 30.4% to 93.7%. Positive predictive value ranged from 33.3% to 42.1%, and negative predictive value ranged from 65.4% to 97.0%. The highest sensitivity for melanoma diagnosis was observed for an application that sends the image directly to a board-certified dermatologist for analysis, and the lowest sensitivity was observed for applications that use automated algorithms to analyze images.	Not provided	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Xiong et al., 2018[36]	Convolutional neural networks model to detect acid-fast stained tuberculosis bacillus	Evaluation of model performance using 246 samples of both positive and negative cases (45 in training set, 201 cas.es in testing set) collected from January 2016 to June 2017	Department of Pathology, Peking University First Hospital	The model achieved a high (97.94%) sensitivity and moderate (83.65%) specificity.	Not provided	Low	None

Table B.2: Diagnostic Errors, Clinical Decision Support—Systematic Reviews and Meta-Analyses

Note: Full references are available in the Section 1.1 reference list.

Author, Year	Description of Patient Safety Practice	Settings and Population	Summary of Findings	Comments
el-Kareh et al., 2013[1]	Diagnostic decision support systems and diagnosis-related health information technology (HIT)	Systematic review of HIT to reduce diagnostic error. The search strategy did not include limitations for settings or populations.	The use of HIT in diagnosis is still in its early stages. Many aspects of the diagnostic process have been targeted, but few tools and systems have been shown to improve diagnosis in actual clinical settings.	Included in Riches, 2016, systematic review and meta-analysis
Graber et al., 2012[7]	Interventions to prevent, reduce, or mitigate diagnostic errors, including CDS to support and improve cognition	Systematic review of cognitive interventions to reduce diagnostic error. The search strategy did not include limitations for settings or populations.	ISABEL has good sensitivity in both pediatric and adult settings, with sensitivity in the adult setting approaching 100%. Research on the use of Google searches yields the correct diagnosis in only 58% of difficult cases.	None
Nurek et al., 2015[40]	Computerized diagnostic decision support systems (CDDSS)	Meta-review of existing systematic reviews of CDS systems in primary care to improve diagnosis. Subjects (primary end-users of CDS) include individual clinicians; no specific criteria for setting.	Identified the following requirements for successful integration of a CDS: a more standardized computable approach to knowledge representation is needed, one that can be readily updated as new knowledge is gained, and a deep integration with the EHR is needed in order to trigger at appropriate points in cognitive workflow.	None
Riches et al., 2016[8]	Differential diagnosis (DDX) generators	Systematic review and meta-analysis investigate the efficacy and utility of DDX generators. Subjects include the individual user of the tool and the clinical case being entered into the tool; no specific criteria for setting.	The pooled accurate diagnosis retrieval rate of DDX tools was high, with high heterogeneity (pooled rate=0.70, 95% CI, 0.63 to 0.77; I2=97%, p<0.0001). DDX generators did not demonstrate improved diagnostic retrieval compared with clinicians, but small improvements were seen in the before and after studies, in which clinicians had the opportunity to revisit their diagnoses following DDX generator consultation.	None

Author, Year	Description of Patient Safety Practice	Settings and Population	Summary of Findings	Comments
Wagholikar et al., 2012[21]	Computer-assisted diagnosis models	Systematic review of modeling techniques to provide diagnostic support. The search strategy did not include limitations for settings or populations. (Search was focused on models.)	General trends in research of medical decision support: • Improvement in the accuracy of MDS application may be possible by modeling of vague and temporal data, research on inference algorithms, integration of patient information from diverse sources, and improvement in gene profiling algorithms. • Research would be facilitated by public release of de-identified medical datasets and development of open-source data-mining tool kits. • Comparative evaluations of different modeling techniques are required to understand characteristics of the techniques and to guide developers in the choice of technique for a particular medical decision problem. • Evaluations of MDS applications in the clinical setting are necessary to foster physicians' use of these decision aids.	None

Table B.3: Diagnostic Errors, Result Notification Systems—Single Studies

Note: Full references are available in the Section 1.2 reference list.

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Chen et al., 2011[21]	Automated phone alert using short message service (SMS)	Pre/post design; total of 223 patients with acid-fast bacilli-positive tuberculosis (96 baseline, 127 post-intervention).	1,600-bed academic medical center, Taiwan	The laboratory delay (p<.001), response delay (p=.045), and interval from admission to transfer to the isolation room (P<.001) were all significantly reduced during the intervention phase. The proportion of patients transferred to isolation within 1 day increased significantly.	Not provided	Need adequate staffing levels to support the RNS and operational changes.	Low	None
Dalal et al., 2014[12]	Automated email system	Cluster-randomized controlled trial; 441 adult general medicine and cardiology patients who had one or more tests pending at discharge (TPAD) and their 117 attending physicians (241 patients/59 attending physicians in intervention arm, 200 patients/58 attending physicians in control arm).	Academic medical center: 720-bed tertiary-care hospital and academic medical center and primary care outpatient setting, United States	There was a statistically significant increase in the rate of awareness of TPAD results by attending physicians for patients assigned to the intervention compared with usual care (76% vs. 38%, adjusted/clustered odds ratio [OR] 6.30, 95% confidence interval [CI], 3.02 to 13.16, p<0.001).	Not provided	Need for connectivity between hospitals and primary care physicians (PCPs) outside of network. Integrate RNS into workflow.	High	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Dalal et al., 2018[13]	Automated email system	Cluster-randomized controlled trial. Attendings and PCPs caring for adult patients discharged from general medicine and cardiology services with at least one actionable TPAD between June 2011 and May 2012; 3,378 TPADs representing 1,522 patient discharges sampled.	Academic medical center: 720-bed tertiary-care hospital and primary care outpatient setting, United States	The proportion of actionable TPADs with documented action was 60.7% vs. 56.3% (p=0.82) in the intervention vs. usual care groups; similar for documented acknowledgment. Pathology tests were the type most commonly associated with documented followup.	Not provided	Need connectivity between hospitals and PCPs outside of network.	Moderate	None
Eisenberg et al., 2010[19]	Manual, web-based electronic messaging system	Post-intervention; 908,475 imaging exams performed, with 10,510 level 3 alerts (abnormal conditions that could result in considerable morbidity if they are not appropriately treated, but which are not immediately life-threatening) submitted to messaging system. Five hundred randomly selected alerts reviewed.	Single large academic medical center with several off-campus outpatient facilities, United States	All results were communicated to the referring providers, with 411 of 500 (82.2% +- 3.3) communications accomplished within the 48-hour policy goal. Note that day of week affected outcome, with more alerts submitted Monday-Thursday before 3 p.m. communicated within 48 hours (93.7% +/- 2.4) than those alerts generated on Thursday afternoon through Sunday (73.0% +/- 9.2).	Not provided	Need adequate staffing to support the RNS. Establish policies and procedures around RNS use.	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
El-Kareh et al., 2012[27]	Automated email system	Cluster-randomized controlled trial; 157 results for 121 total inpatient and outpatient physicians (73 in intervention group, 48 in control group) caring for hospitalized adult patients with positive and untreated/ undertreated culture results returned after discharge.	Academic hospital (777 beds) and primary care outpatient settings; United States	Twenty-seven out of 97 (28%) results in the intervention group and 8 out of 60 (13%) in the control group (aOR 3.2, 95% CI, 1.3 to 8.4; p=0.01) had documented followup in the outpatient chart within 3 days of post-discharge result.	Not provided	Integrate RNS into workflow.	Low	None
Etchells et al., 2010[10]	Automated paging system	Randomized controlled trial; 165 critical lab values with documented response time (81 intervention; 84 control) on 108 patients admitted to the four general medicine clinical teaching units.	General medicine clinical teaching units at an urban academic hospital, Canada	There was a 23-minute reduction in median response time (interval between acceptance of the critical value into the LIS and the documented writing of order or documented time of treatment), but this was not statistically significant. Median response time was 16 min (IQR 2-141) for the automated paging group and 39.5 min (IQR 7-104.5) for the usual care group (p= 0.33).	Some critical results, such as those from repeated troponin tests, were viewed as nuisances. The physician-on-call had to carry numerous additional pagers and could not always discern which pager was alerting.	Automated physician scheduling integrated with RNS. Establish policies and procedures around RNS use.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Etchells et al., 2011[11]	Automated alerts via mobile phone or pager and link to CDS for alert	Randomized controlled trial (controlled stepped-wedge design); general internal medicine teaching units; 498 critical laboratory conditions on 271 patients.	General internal medicine clinical teaching units at two academic hospitals, Canada	Overall, 50% of potential clinical actions were carried out, and there were adverse clinical events within 48 hours for 36% of the laboratory conditions. The median (IQR) proportion of potential clinical actions that were actually completed was 50% (33%–75%) with alerting system on, and 50% (33–100%) with alerting system off (p=0.94, Wilcoxon rank sum test). When the alerting system was on (n=164 alerts) there were 67 adverse events within 48 hours of the alerts (42%). When the alerting system was off (n=334 alerts), there were 112 adverse events within 48 hours (33%; difference: 9% higher number of adverse events with alerting system on, p=0.06).	Not provided	Automated physician scheduling integrated with RNS. Establish policies and procedures around RNS use.	Low	None
Lacson et al., 2014[15]	Manual-triggered alert via pager or email	Pre/post design; 47,034 reports randomly sampled and manually reviewed (9,430 1 year prior to intervention; 37,604 4 years post-intervention).	Academic medical center (753 beds), United States	Adherence to the institutional policy for timely closed-loop communication of critical imaging results increased from 91.3% before the intervention to 95.0% after the intervention (p<0.0001). There was a ninefold increase in the critical results communicated via the system (chi-square trend test, p<0.0001).	Not provided	Establish policies and procedures around RNS use.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Lacson et al., 2016[16]	Manual-triggered alert via pager or email	Trend analysis; 10 semi-annual time periods from 42 randomly selected radiology reports from each of 10 semi-annual time periods between 2009 and 2014; total of 840 reports, 420 with documented communication and 420 without documented communication.	Single adult quaternary referral academic medical center, United States	After the implementation of the critical imaging test result policy and the ANCR, critical results lacking documented communication decreased nearly fourfold between 2009 and 2014 (0.19 to 0.05, p<0.0001).	There was concern over alert fatigue, a potential unintended consequence of implementing alerting systems, but authors did not find an increase in non-clinically significant results communicated through the system.	Establish policies and procedures around RNS use. Integrate RNS into workflow.	Low	None
Lin et al., 2014[23]	Automated phone text-message alert	Pre/post design. Patients with warfarin therapy managed by the hospital's outpatient clinics; 3,497 patients (30,981 tests) were included in the manual alert study period and 3,781 patients (32,297 tests) were included in the PHS alert group.	Outpatient department of a 2,500-bed tertiary teaching hospital, Taiwan	Incidence of major thromboembolic events was 1.6% pre-intervention and 1.6% post-intervention (p=0.709), and the rate of hemorrhagic events was 3.1% and 4.2% in the manual alert and PHS alert study periods (p=0.198).	Not provided	In hospital, need RNS technology to be available to all stakeholders.	Low	Study examines patient outcomes

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
O'Connor et al., 2016[17]	Manually triggered alert via pager or email/alert in electronic medical record (EMR)	Pre/post design; 171 PCPs at 13 affiliated outpatient practices; 5,931 outpatient nonurgent, clinically significant radiology alerts (1,503 pre-intervention; 4,428 post-intervention).	Tertiary academic medical center (793 beds) and affiliated outpatient practices, United States	There was 100% acknowledgement of non-urgent clinically significant ANCR-generated alerts, with the EHR used to acknowledge 15.5% of them. Ninety percent of alerts pre-intervention and 84% post-intervention were actionable (p=.29). PCPs acted on 94% (85 of 90; 95% CI, 88 to 98) of actionable alerts pre-intervention and 94% (79 of 84; 95% CI, 87 to 97) post intervention (p>.99).	Not provided	Integrate the RNS into workflow. Establish policies and procedures around RNS use.	Low	None
O'Connor et al., 2018[24]	Manually triggered alert via pager or email	Pre/post design; 5,595 pathology reports with malignancies (2,793 pre-intervention; 2,802 post-intervention).	Community hospital (150 beds) affiliated with an academic medical center, United States	Acknowledgment of the CSTR within 15 days, the institutional policy, was documented for 98 of 107 (91.6%) pre-intervention reports and 89 of 103 (86.4%) post-intervention reports (p=0.2294). Median time to acknowledgment was 7 days (interquartile range [IQR], 3, 11) pre-intervention and 6 days (IQR, 2, 10) post-intervention (p=0.5083). Post-intervention, median time to acknowledgment was 2 days (IQR, 1, 6) for reports with ANCR alerts versus 6 days (IQR, 2.75, 9) for reports without alerts (p=0.0351).	Not provided	Provide review and feedback about use of RNS. Establish policies and procedures for RNS use.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Park et al., 2008[20]	Automated phone alert using SMS and callback	Pre/post design; 217 critical hyperkalemia alerts (121 pre-intervention; 96 post-intervention).	Tertiary care academic medical center (2,200 beds), South Korea	Across all wards (intensive care units [ICUs] and general wards), the median and interquartile ranges of the clinical response times were significantly reduced, going from 213.0 min and 476.0 min to 74.5 min and 241 min, respectively (p<.001). The mean and median clinical response times in general wards were significantly decreased by 54.3% and 74.7%, respectively, in comparison to the pre-intervention response times (p<.001). The mean and median clinical response times in ICUs decreased by 11.8% and 51.8%, respectively, in comparison to those in 2001, but the change was not significant (p=.190).	Not provided	Need to account for technology limitations such as inconsistent phone reception within the hospitals.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Singh et al., 2009[18]	Automated EMR alert notification system	Post-intervention; 123,638 radiology studies generating 1,196 alerts, of which 979 (81.9%) were tracked as acknowledged and 217 (18.1%) were unacknowledged.	Single multi-specialty ambulatory clinic and five satellite clinics affiliated with U.S. Department of Veterans Affairs (VA), United States	Nine hundred seventy-nine (81.9%) alerts were tracked as acknowledged and 217 (18.1%) were unacknowledged. For 131 (11%) of alerts, there was no evidence of documented followup. There were 92 (7.7%) results without timely followup at 4 weeks after result transmission. Lack of acknowledgement was associated with physician assistants as ordering providers compared with attending physicians (OR: 0.46; 95% CI, 0.22 to 0.98), trainees as ordering providers (OR: 5.58; 95% CI, 2.86 to 10.89), and when dual as opposed to single communication was used (OR: 2.02; 95% CI, 1.22 to 3.36). There was no significant difference in rates of lack of timely followup between the acknowledged and unacknowledged alerts (7.3% vs. 9.7%; p=0.2).	Dual communication, intended to be a “safeguard” to protect against loss of followup, was unexpectedly associated with lack of timely followup.	Establish policies and procedures for RNS use.	Low	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Singh et al., 2010[5]	Automated EMR alert notification system	Observational/cross-sectional; 78,158 laboratory tests (HbA1c, Hep B Ab, PSA, TSH) performed, with 1,163 results transmitted as mandatory high-priority alerts (1.49% of results screened).	Single multispecialty ambulatory clinic and five satellite clinics affiliated with VA, United States	Of the alerts, 6.8% lacked timely followup at 30 days. Lack of acknowledgement was associated with allied health care providers as ordering providers (OR, 4.32; 95% CI, 1.21 to 15.52) and trainees as ordering providers (OR, 8.39, 95% CI, 2.97 to 23.68), compared with attending physicians. Specialty services were found less likely to acknowledge alerts compared with primary care providers (p<.0001). There was no significant difference in rates of lack of timely followup between acknowledged and unacknowledged laboratory alerts (6.4% vs. 10.1%; p=.13) and no significant differences in ordering provider types (p=.67), but there was a significant difference across specialties (p<.0001).	Not provided	Not provided	Low	None

Table B.4: Diagnostic Errors, Result Notification Systems—Systematic Reviews and Meta-Analyses

Note: Full references are available in the Section 1.2 reference list.

Author, Year	Description of Patient Safety Practice	Settings and Population	Summary of Findings	Implementation Themes/Findings
Liebow et al., 2012[8]	Automated notification systems; call centers	Nine articles met criteria for inclusion, as follows. Population: All patients in healthcare settings with lab results that include a critical value. Intervention: Automated notification systems and call centers for communicating critical values. Comparison: Manual critical values notification systems. Outcome: Timeliness and accuracy of reporting or receipt of critical values information, or timeliness of treatment based on critical values information.	Automatic notification systems (4 studies): only one study of "good quality"; average improvement from implementing automated notification systems is d=0.42 (95% confidence interval [CI], 0.2 to 0.62). Overall strength of evidence is suggestive. Call centers (5 studies): the average odds ratio for call centers is odds ratio [OR]=22.1 (95% CI, 17.1 to 28.6). Call centers are effective in improving the timeliness and accuracy of critical value reporting in an inpatient care setting, and are recommended as an "evidence-based best practice."	Automated notification systems may disrupt usual lines of communication and provide too much/too frequent information. Risk of losing back-up contact information; risk for HIPAA violations. Call centers may require additional communications with lab staff when caregivers require additional information that call centers may not have; staffing needs are significant.
Slovis et al., 2017[9]	Automated notification systems (asynchronous)	Thirty-four articles pertaining to asynchronous automated electronic notifications of laboratory results published through 2016.	Several asynchronous automated electronic notification systems for laboratory results have been successfully implemented with improvements in workflow and time to acknowledgement of results.	Though some critical alerts are necessary, not all critical results warrant notification, because not all critically abnormal laboratory values require emergent intervention. However, some studies have demonstrated that noncritical urgent and elective notifications can also improve clinical care.

Table B.5: Diagnostic Errors, Education and Training—Single Studies

Note: Full references are located in the Section 1.3 reference list.

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)
Coderre et al., 2010[20]	Use of querying an initial hypothesis to generate cognitive reflection in medical students	Pre/post study design with comparison groups; 67 first-year medical students	University of Calgary, Canada	Questioning an initial diagnosis through processing of additional data does not affect a correct initial diagnosis, but it does allow correction of an inaccurate initial diagnosis.	Not provided	Moderate
Dyre et al., 2017[38]	Error management training (2 components: active exploration during skill practice and the provision of error management instructions)	Randomized trial; medical students with no prior ultrasound experience; 32 students received error management training (EMT) and 28 received error avoidance training (EAT)	Department of Obstetrics, Rigshospitalet, Denmark	Providing error management instructions, rather than error-avoidance instructions, during simulation-based training improved the transfer of learning to the clinical setting. Mean test scores in the transfer test corresponded to a large effect size in favor of EMT (Cohen's d=1.11, 95% confidence interval [CI], 0.5 to 1.7).	Not provided	Low
Goodman and Kelleher, 2017[32]	Focused session of interpretation training at a local art gallery where art experts taught the trainees how to thoroughly analyze a painting	Pre/post study design, no comparison group; 15 first-year radiology residents	Not provided	Focused teaching on perception improved first-year residents' ability to localize imaging abnormalities. For the pretest, residents scored an average of 2.3 out of a maximum possible score of 15 (standard deviation (SD) of 1.4, range of 0–4). After training, average post-test score increased to 6.3 (SD of 1.8, range of 3–9) ($p < .0001$).	Not provided	Moderate
Mamede et al., 2010[17]	Structured reflection as taught through the use of five steps aimed at inducing reflective reasoning	Pre/post study design, with comparison group; 18 first-year and 18 second-year internal medicine residents	Erasmus Medical Centre, Rotterdam, Netherlands	When establishing diagnoses using nonanalytic reasoning, availability bias may occur in response to recent experience with similar cases. This bias may be counteracted by using reflective reasoning. Reflection improved all participants' diagnoses compared with nonanalytical reasoning.	Reflective practice may take its full effect only with more difficult clinical scenarios.	Low to moderate

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)
Mamede et al., 2012[18]	Compared structured reflection with providing a single diagnosis or generating differential diagnoses while practicing clinical cases	Three-phase experimental study; 46 fourth-year medical students	Erasmus Medical Centre, Rotterdam, Netherlands	Using structured reflection to diagnose cases increases the learning of clinical knowledge more effectively than using immediate diagnosis or differential diagnosis generation.	Not provided	Low to moderate
Mamede et al., 2014[19]	Compared structured reflection with providing a single diagnosis or generating differential diagnoses while practicing clinical cases	Two-phase experimental study; 110 fourth-year medical students	Erasmus Medical Centre, Rotterdam, Netherlands	Use of structured reflection was more effective in supporting learning than providing a single diagnosis or differential diagnoses.	Not provided	Low to moderate
McFadden and Crim, 2016[41]	Online simulation-based training activity to improve diagnosis; training supplemented with interactive practice opportunities and feedback delivered by an artificial intelligence–driven simulation/tutor	Pre/post design with comparison group using convenience sampling; 68 practicing primary care practitioners (27 in control group, 41 in treatment group)	Continuing medical education (CME) conference (control group), standalone online CME (intervention group)	There was no difference between control and intervention groups in pre-training diagnostic accuracy. The control group's post-training performance did not statistically significantly improve (p=.13); the intervention group's post-training diagnostic performance significantly improved, by 22% (p<.02).	Not provided	Low
Mohan et al., 2018[26]	Virtual simulation using two "serious" video games to train on the use of a heuristic, judgment by representativeness	A randomized controlled trial, using 257 board-eligible or board-certified emergency medicine physicians who worked primarily at non-trauma or level III/IV trauma centers	American College of Emergency Physicians Scientific Assembly	Both game interventions reduced under-triage events on the simulation compared with the control condition, whereas the text-based intervention did not.	Not provided	Low
Nendaz et al., 2011[21]	Weekly in-person case-based clinical reasoning seminars incorporating diagnostic reflection	Randomized controlled study; 29 medical students (14 in the control group and 15 in the intervention group, providing 28 and 30 encounters, respectively)	University of Geneva Faculty of Medicine, Switzerland	The case-based clinical reasoning seminars did not significantly affect the students' overall diagnostic or decisional competencies, but did aid in increasing the relevance of their differential diagnoses as written in the post-encounter notes.	Reflective practice may take its full effect only with more-difficult clinical scenarios.	Moderate

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)
Pusic et al., 2012[35]	Radiographic training sets, which varied in their proportions of abnormal cases (30%, 50%, 70%)	Prospective, double-blind, randomized, three-arm education trial; 100 residents completed the study	Six academic training programs for emergency medicine and pediatric residents, United States	The two groups did not differ in accuracy on the post-test (p=0.20). The group with a low proportion of abnormal cases had the highest false negative rate, and missed fractures one-third more often than the groups that trained on higher proportions of abnormal cases. Manipulating the ratio of abnormal to normal cases the students are exposed to can alter their sensitivity and specificity.	Online educational intervention	Low
Reilly et al., 2013[22]	Three-part, 1-year curriculum in cognitive bias and diagnostic error	Pre/post study design with comparison group; 38 PGY-2 internal medicine residents	Perelman School of Medicine at the University of Pennsylvania, United States	Performance on the 13-item multiple-choice knowledge test improved post-curriculum when compared with both pre-curriculum performance (9.26 vs. 8.26, p=0.002) and the PGY-3 comparator group (9.26 vs. 7.69, p<0.001). Residents who participated in this curriculum improved their recognition and knowledge of common cognitive biases and heuristics.	Not provided	Moderate
Schwartz et al., 2010[39]	Four weekly case-based 1-hour in-person didactic sessions to help the students develop knowledge and skills in contextualizing patient care	Quasi-randomized controlled trial; 124 fourth-year medical students in internal medicine sub-internships	University of Illinois at Chicago and Jesse Brown Veterans Administration Medical Center, United States	Students who participated in the contextualization workshops were significantly more likely to probe for contextual issues in the standardized patient encounters than students who did not, and significantly more likely to develop appropriate treatment plans for standardized patients with contextual issues.	Not provided	Moderate

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)
Sherbino et al., 2011[14]	A 90-minute, standardized, interactive, case-based teaching seminar on cognitive forcing strategies (CFS)	Cross-over study design; consecutive enrollment of 56 senior medical students during their emergency medicine rotation	McMaster University	Preliminary findings suggest that application of CFS and retention are poor. Even immediately after instruction, in a test situation that is deliberately linked to the educational intervention, fewer than half the students in the study used CFS to correctly "de-bias" themselves. Two weeks post-CFS training, there was no evidence of de-biasing.	Not provided	Moderate
Sherbino et al., 2014[15]	A 90-minute, standardized, interactive, case-based teaching seminar on CFS	Prospective, controlled trial; 198 senior medical students in EM rotation (145 in intervention, 46 in control group)	McMaster University	The educational interventions employed to teach CFS failed to show any reduction in diagnostic error by novices.	Not provided	Low to moderate
Smith et al., 2009[40]	Four-month online didactic continuing education program to improve ability of rural radiographers to interpret plain musculoskeletal radiographic examinations	Pre/post design, no comparison group; 16 rural radiographers	Northern Sector of the Hunter New England Area Health Service, UK	Short-term intensive training can improve diagnostic accuracy of rural radiographers. There was a statistically significant improvement at the "general opinion" and "observation" levels for the more complex cases (paired t-test, $p<0.05$), while there was no change in image interpretation accuracy for less complex cases.	Online educational intervention	Moderate
Smith and Slack, 2015[16]	Workshop on debiasing (taught to recognize and respond to cognitive biases), including training reflective exercises	Pre/post study, no comparison group; 19 family medicine residents	Family Medicine Residency Program at David Grant Medical Center, Travis Air Force Base, California, United States	After the workshop, residents' formulation of an acceptable plan to mitigate the effect of cognitive bias increased from 84% (36 of 43) to 100% (33 of 33, $p=0.02$). There was no effect on preceptor concurrence with the residents' diagnoses, the residents' ability to recognize their risk of cognitive bias, or the preceptors' perception of an unrecognized cognitive bias in the residents' presentation.	Not provided	Moderate

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)
Soh et al., 2013[34]	One-hour online e-learning tutorial to improve visual perception skills	Randomized controlled trial, 14 first-year medical radiation sciences students (technologists)	Medical radiation science program, Australia	The experiment group demonstrated a 45% increase in the mean number of fixations per case (p=.047), with a 30% increase in sensitivity (p=.022), following the tutorial. The experiment group also demonstrated improved lesion detection overall and a 49% decrease in mean time to first fixation on the lesion (p=.016).	Online educational intervention	Moderate
van der Gijp et al., 2017[33]	Training on two visual search strategies, “scanning” and “drilling,” used in radiology to improve visual perception	Randomized cross-over design; 19 first- and second-year radiology residents	Academic medical center’s radiology residency program, United States	Perceptual performance following drilling search instructions outperformed performance following scanning search instruction in terms of true positives.	Not provided	Moderate
Wolpaw et al., 2009[9]	Training on the use of SNAPPS technique—Summarize history and findings, Narrow the differential; Analyze the differential; Probe preceptor about uncertainties; Plan management; Select case-related issues for self-study—for case presentations to facilitate learning	Post-test-only, comparison groups, randomized trial; 108 third-year medical students	Case Western Reserve University School of Medicine, United States	SNAPPS group showed more diagnostic reasoning than a feedback comparison and a control group.	Not provided	Moderate (qualitative analysis)

Table B.6: Diagnostic Errors, Education and Training—Systematic Reviews and Meta-Analyses

Note: Full references are available in the Section 1.3 reference list.

Author, Year	Description of Patient Safety Practice	Settings and Population	Summary of Findings
Cook et al., 2010[7]	Virtual patients	Studies published in any language that investigated use of a virtual patient to teach health professions learners. Virtual patient is "a specific type of computer program that simulates real-life clinical scenarios; learners emulate the roles of healthcare providers to obtain a history, conduct a physical exam, and make diagnostic and therapeutic decisions." No beginning date cutoff, and the last date of search was February 16, 2009.	Systematic review and meta-analyses. Included 4 qualitative studies, 18 no-intervention controlled studies, 21 noncomputer instruction comparative studies, and 11 computer-assisted instruction comparative studies. Use of virtual patients was associated with large positive effects compared with no intervention.
Graber et al., 2012[8]	Various interventions, including educational interventions	Articles and books that contained results from an intervention trial or suggested an intervention to reduce cognitive-related diagnostic error.	Review included 141 sources (42 empirical studies; 100 contained suggestions for interventions; and 1 had both). The review focused on three areas to reduce diagnostic errors: increase knowledge and experience, improve clinical reasoning, and get help.
McDonald and Matesic, 2013[36]	Patient safety strategies targeting diagnostic errors, including educational interventions	Studies that evaluated any intervention to decrease diagnostic errors (incorrect diagnoses or missed diagnoses) in any clinical setting and with any study design and patient outcomes.	Eleven studies used educational interventions aimed at various populations. Strategies targeted at clinicians produced improvements, but the studies were nonrandomized. Two randomized trials that targeted consumers in the diagnostic process found improvements.

Table B.7: Diagnostic Errors, Peer Review—Single Studies

Note: Full references are available in the Section 1.4 reference list.

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Agrawal et al., 2017[18]	Simultaneous double-reporting of emergency teleradiology examinations with discrepancies adjudicated by the radiologists before finalization of the report	Descriptive analysis of retrospective data; 3,779 double-read radiological procedures over 4 months	International teleradiology practice and two non-teaching mid-sized to large community hospitals, United States	Of the 145/3,779 procedures (3.8%; 95% confidence interval [CI], 3.2 to 4.4) for which the double-reporting identified undetected or incompletely evaluated findings that led to report modifications, 69 were clinically significant. MRI spine studies contributed significantly more than other study types to these errors.	Not provided	To promote efficiency, limit double reviews to certain study types that have the greatest risk of diagnostic errors.	Moderate	In Geijer, 2018
Harvey et al., 2016[10]	Regularly scheduled consensus-oriented group reviews (3 or more radiologists) of randomly selected recently interpreted computerized tomography (CT), magnetic resonance imaging (MRI), and ultrasound cases (within 3–7 days)	Descriptive analysis of retrospective data. A total of 11,222 studies reported by 83 radiologists were peer-reviewed using COGR at 2,027 conferences during the 2-year study period	Radiology department at a 950-bed tertiary care academic center, United States	The average radiologist participated in 112 peer review conferences and had 3.3% of their available CT, MRI, and ultrasound studies peer reviewed. The discordance rate was 2.7% (95% CI, 2.4 to 3.0), with significant differences found on the basis of division and modality.	Not provided	Necessary to have stakeholder buy-in. Implementation associated with increased staffing needs, workload, and associated costs. Concern over maintenance of confidentiality may affect implementation.	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Itri et al., 2018[11]	Peer review of randomly selected (20 cases/month adjudicated by third party) and nonrandomly selected (diagnostic errors found during routine clinical practice) radiologist interpretations and peer learning conferences (PLCs)	Descriptive analysis of retrospective data; 1,880 total abdominal imaging cases (190 identified via nonrandom peer review process; 1,690 identified via random peer review process) read by 10 radiologists	Abdominal imaging section of a radiology department in an academic tertiary care medical center, United States	Random peer review process: 1,690 cases reviewed, 2.6% with incidental errors. None considered to be significant or major discrepancies. Nonrandom process: 190 cases identified, 94 categorized as significant, 36 categorized as major discrepancies. CTs and MRIs accounted for 164 of the cases.	Not provided	Not provided	Moderate	None
Kamat et al., 2011[15]	Laboratory information system-driven pre-signout quality assurance tool to randomly select an adjustable percentage of pathology cases for peer review and adjudication by the pathologists prior to release of the final report	Descriptive analysis of retrospective data; 1,339 (7.45%) out of a total 17,967 non-gynecologic cytopathology cases over an 18-month period	Pathology department at a university medical center, United States	In 2.6% of cases there were discrepancies, including 34 minor and 1 major.	Not provided	Implementation associated with increased staffing needs, workload, and associated costs.	Moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Lauritzen, 2016[19]	Prospective radiologist-requested double-reading of CT abdomen examinations	Retrospective cross-sectional study; 1,071 consecutive double-reported abdominal CT examinations of surgical patients	Multicenter study; five public hospitals, Norway	Of 1,071 reports, 146 contained clinically important changes (14%, 95% CI, 11.6 to 15.8), with changes to 108 reports (10%, 95% CI, 8.3 to 12.0) considered intermediate, 35 major (3%, 95% CI, 2.3 to 4.5), and 3 critical (0.3%, 95% CI, 0.06 to 0.8).	Not provided	Concern over maintenance of confidentiality may affect implementation.	Low to moderate	In Geijer, 2018
Lauritzen et al., 2016[20]	Prospective radiologist-requested double-reading of CT chest examinations	Retrospective cross-sectional study; 1,023 consecutive double-reported chest CT examinations	Multicenter study; five public hospitals, Norway	Report changes were classified as clinically important in 91 (9%) of 1,023 reports. Of these, 3 were critical (demanding immediate action), 15 were major (implying a change in treatment), and 73 were intermediate (affecting subsequent investigations).	Not provided	Not provided	Low to moderate	In Geijer, 2018

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Layfield and Frazier, 2017[14]	Random peer review (10% of all surgical pathology cases); nonrandom peer review (solicited review correlation of internal and external diagnoses; unsolicited correlation of internal and external diagnoses in cases sent for review at a second institution treating the patient; and review of all dermatopathology cases)	Descriptive analysis of retrospective data; all cases undergoing review by any of the four review protocols over a 1-year period were included	Department of Pathology and Anatomical Sciences at a university medical center, United States	The 10% random review detected 17 errors in 2,147 cases (0.8%); solicited case consultations detected 5 errors in 70 cases (7.1%); unsolicited reviews by outside institutions detected 3 errors in 190 cases (1.6%); and focused reviews of dermatopathology cases identified 5 errors in 59 cases (8.5%).	Not provided	Implementation associated with increased staffing needs, workload, and associated costs.	Moderate	None
Lian et al., 2011[22]	Retrospective review by two subspecialists of initially double-read CT angiography studies (head and neck); initial studies read by a staff neuroradiologist alone, by staff and diagnostic radiology resident, and by staff and neuroradiology fellow	Descriptive analysis of retrospective data; 503 sequential neck and intracranial CTA studies performed over a 6-month period	Unspecified	Reviewed 503 studies; 144 were originally reported by a staff neuroradiologist alone, 209 by staff and a diagnostic radiology resident, and 150 by staff and a neuroradiology fellow. Twenty-six significant discrepancies were discovered in 20/503 studies (4.0%).	Not provided	Not provided	Moderate	In Geijer, 2018

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Lindgren et al., 2014[25]	Retrospective interpretations of radiology studies (CT, MRI, and ultrasound abdominal studies) initially performed at an outside institution	Descriptive analysis of retrospective data; 398 abdominal imaging reinterpretations performed on 380 patients between 1/1/2010 and 7/15/2010	Single hospital, United States	Three hundred ninety-eight report comparisons were reviewed on 380 patients. The initial report had 5.0% (20/398) high clinical impact interpretive discrepancies and 7.5% (30/398) medium clinical impact discrepancies. The subspecialized secondary report had no high clinical impact discrepancies and 8/398 (2.0%) medium clinical impact discrepancies.	Not provided	Not provided	Moderate	In Geijer, 2018
Murphy et al., 2010[21]	Prospective, blinded double-reporting of minimal-preparation CT colon (MPCTC) with discrepancies resolved by followup colonoscopies	Prospective cohort of 186 consecutive patients undergoing MPCTC for lower gastrointestinal symptoms	Single hospital; UK	Of the 186 imaging reports, 111 had at least one discrepancy (60%). Sixty-seven clinically relevant extracolonic lesions were identified (25 identified in one report, 42 in both), and 24 clinically relevant colonic lesions (7 in one report, 17 in both). Of the 17 colonic lesions reported by both radiologists, 5 were false positives as determined by normal colonoscopies. Of, the 7 reported by one radiologist, 1 was a biopsy-proved cancer.	Increased false-positives. Double-reporting found one extra-colonic cancer, but at the expense of five unnecessary endoscopic procedures.	Implementation associated with increased staffing needs, workload, and associated costs.	Low	In Geijer, 2018

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Natarajan et al., 2017[23]	Retrospective reinterpretations by radiologists of plain radiographs initially read by pediatric orthopedists	Retrospective cohort; 1,570 consecutive pediatric orthopedic clinic patients with 2,509 radiographic studies during a 4-month period	Pediatric orthopedic clinic in an academic children's hospital, United States	Of 2,264 radiographic studies reviewed by a radiologist, new, clinically important information was added in 23 (1.0%) of studies. In 38 (1.7%) of the studies, the radiologist review missed the diagnosis or clinically important information that could affect treatment.	Not provided	Implementation associated with increased staffing needs, workload, and associated costs.	Low to moderate	In Geijer, 2018
Onwubiko and Mooney, 2016[24]	Retrospective reinterpretations of pediatric trauma CT scans initially performed at outside institution	Descriptive analysis of retrospective data; 168 patients transferred with CT abdomen and pelvis scans performed at outside institutions	Level 1 pediatric trauma center, United States	Ninety-eight CT abdomen/pelvis scans were reinterpreted, with 12 new, clinically significant injuries detected. Three patients had solid organ injuries upgraded and four were downgraded to no injury.	Not provided	Implementation associated with increased staffing needs, workload, and associated costs.	Low to moderate	In Geijer, 2018
Raab et al., 2008[12]	Random peer review (5% of cases) and focused secondary review (known diagnostically challenging case types) of surgical pathology cases	Nonconcurrent cohort study; 7,444 cases from random review process and 380 cases reviewed using focused review process	Single site within a large multihospital system, United States	The numbers of errors detected by the targeted 5% random and focused review processes were 195 (2.6% of reviewed cases) and 50 (13.2%), respectively (p<.001). The numbers of major errors for the targeted 5% random and focused review processes were 27 (0.36%) and 12 (3.2%), respectively (p<.001).	Not provided	To promote efficiency, limit double reviews to certain study types that have the greatest risk of diagnostic errors.	Low to moderate	None

Author, Year	Description of Patient Safety Practice	Study Design; Sample Size; Patient Population	Setting	Outcomes: Benefits	Outcomes: Harms	Implementation Themes/ Findings	Risk of Bias (High, Moderate, Low)	Comments
Swanson et al., 2012[13]	Peer review of randomly selected radiology studies (4 cases/shift) and voluntary, nonrandom case review with feedback	Descriptive analysis; peer review reports on 5,278 radiologic studies (4,892 mandatory random review; 386 voluntary review) conducted over 4-year period	Large urban multidisciplinary children's hospital, United States	The discrepancy rate was 3.6% between original interpretation and random peer review and 12% for the nonrandom review.	Not provided	Not provided	Moderate	None

Table B.8: Diagnostic Errors, Peer Review—Systematic Reviews and Meta-Analyses

Note: Full references are available in the Section 1.4 reference list.

Author, Year	Description of Patient Safety Practice	Settings and Population	Summary of Findings	Implementation Themes/Findings
Geijer and Geijer, 2018[16]	Double-reading of radiology studies	Included studies calculating the rate of misses and overcalls with the aim of establishing the added value of double reading by human observers.	Forty-six studies met inclusion criteria. The discrepancy rates varied from 0.4 to 22% in various studies. Double-reading by sub-specialists found high discrepancy rates. Double-reading generally increased sensitivity at the cost of decreased specificity.	To promote efficiency, limit double reviews to certain study types that have the greatest risk of diagnostic errors. Implementation associated with increased staffing needs, workload, and associated costs.
Pow et al., 2016[17]	Double-reading of radiology studies	Studies reporting on the effect of double-reporting on measures of diagnostic efficacy in all imaging modalities, both screening and diagnostic, including sensitivity, specificity, recall rate, and cancer detection rate were included.	Forty-one studies met inclusion criteria. The use of double- reading was found to increase sensitivity and reduce specificity, making it most useful for screening studies where high sensitivity is desired. The authors recommended the use of double-reading in trauma and found that the level of expertise of the reviewers influences the error rate, with those using a subspecialist for the second review having higher rates than for two radiologists with similar training.	To promote efficiency, limit double reviews to certain study types that have the greatest risk of diagnostic errors.

Appendix C. Diagnostic Errors SearchTerms

Method	Search	Search String for: CINAHL	Search String for: MEDLINE
Search 2008-Present, English Only MedLine Publication Types: • Clinical Trial • Clinical Trial, Phase I • Clinical Trial, Phase II • Clinical Trial, Phase III • Clinical Trial, Phase IV • Comparative Study • Controlled Clinical Trial • Corrected and Republished Article • Evaluation Studies • Guideline • Journal Article • Meta-Analysis • Multicenter Study • Practice Guideline • Published Erratum • Randomized Controlled Trial • Review • Scientific Integrity Review • Technical Report • Twin Study • Validation Studies CINAHL Publication Types: • Clinical Trial • Corrected Article • Journal Article • Meta-Analysis • Meta Synthesis	Clinical Decision Support	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Diagnostic Errors" OR "Error(s), Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis")) AND ((MH "Decision Support Systems, Clinical" OR ("Medical Informatics Applications" AND "Information Systems") OR "Reminder Systems" OR "Decision Making, Computer-Assisted" OR "Decision Support Techniques" OR "Diagnosis, Computer-Assisted" OR "Diagnosis, Differential" OR "Artificial Intelligence" AND "Machine Learning" OR "Decision Making, Organizational") OR AB (“Clinical Decision Support” OR (“Medical Informatics Applications” AND “Information Systems”) OR "Decision Support Systems, Clinical" OR "Reminder Systems" OR “Decision Making, Computer-Assisted” OR "Diagnosis, Computer-Assisted" OR “Decision Support Techniques" OR “Artificial Intelligence” OR “IBM Watson” OR “Machine Learning” OR “Decision Support Techniques” OR “Decision Making, Organizational” OR “Differential Diagnosis Generation” OR "Diagnostic Algorithms" OR "Clinical Algorithms" OR "Test Selection Support")))	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Diagnostic Errors" OR "Error(s), Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis")) AND ((MH "Decision Support Systems, Clinical" OR ("Medical Informatics Applications" AND "Information Systems") OR "Reminder Systems" OR "Decision Making, Computer-Assisted" OR "Decision Support Techniques" OR "Diagnosis, Computer-Assisted" OR "Diagnosis, Differential" OR "Artificial Intelligence" AND "Machine Learning" OR "Decision Making, Organizational") OR (AB “Clinical Decision Support” OR (“Medical Informatics Applications” AND “Information Systems”) OR "Decision Support Systems, Clinical" OR "Reminder Systems" OR “Decision Making, Computer-Assisted” OR "Diagnosis, Computer-Assisted" OR “Decision Support Techniques" OR “Artificial Intelligence” OR “IBM Watson” OR “Machine Learning” OR “Decision Support Techniques” OR “Decision Making, Organizational” OR “Differential Diagnosis Generation” OR "Diagnostic Algorithms" OR "Clinical Algorithms" OR "Test Selection Support")))

Method	Search	Search String for: CINAHL	Search String for: MEDLINE
• Practice Guidelines • Randomized Controlled Trial • Research Review • Systematic Review			
Search 2008-Present, English Only MedLine Publication Types: • Clinical Trial • Clinical Trial, Phase I • Clinical Trial, Phase II • Clinical Trial, Phase III • Clinical Trial, Phase IV • Comparative Study • Controlled Clinical Trial • Corrected and Republished Article • Evaluation Studies • Guideline • Journal Article • Meta-Analysis • Multicenter Study • Practice Guideline • Published Erratum • Randomized Controlled Trial • Review • Scientific Integrity Review • Technical Report • Twin Study • Validation Studies	Performance Review and Feedback	(((MH "Diagnostic Errors/PC" OR “Delayed Diagnosis/PC”) OR (AB "Diagnostic Error*" OR "Error*, Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis" OR "Diagnostic Error* Prevention" OR "Diagnostic Error* Control")) AND ((MH "Peer Review" OR "Peer Review, Health Care/ST" OR "Quality Assurance, Health Care/ST" OR "Feedback") OR (AB "Performance Review" OR "Performance Feedback" OR "Clinical Correlation" OR "Peer Review" OR "Feedback" OR "Quality Assurance" OR "Standards" OR "Human Performance")))	(((MH "Diagnostic Errors/PC" OR “Delayed Diagnosis/PC”) OR (AB "Diagnostic Error*" OR "Error*, Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis" OR "Diagnostic Error* Prevention" OR "Diagnostic Error* Control")) AND ((MH "Peer Review" OR "Peer Review, Health Care/ST" OR "Quality Assurance, Health Care/ST" OR "Feedback") OR (AB "Performance Review" OR "Performance Feedback" OR "Clinical Correlation" OR "Peer Review" OR "Feedback" OR "Quality Assurance" OR "Standards" OR "Human Performance")))

Method	**Search**	**Search String for: CINAHL**	**Search String for: MEDLINE**
CINAHL Publication Types: • Clinical Trial • Corrected Article • Journal Article • Meta-Analysis • Meta Synthesis • Practice Guidelines • Randomized Controlled Trial • Research Review • Systematic Review			
Search 2008-Present, English Only MedLine Publication Types: • Clinical Trial • Clinical Trial, Phase I • Clinical Trial, Phase II • Clinical Trial, Phase III • Clinical Trial, Phase IV • Comparative Study • Controlled Clinical Trial • Corrected and Republished Article • Evaluation Studies • Guideline • Journal Article • Meta-Analysis • Multicenter Study • Practice Guideline • Published Erratum • Randomized Controlled Trial • Review	Result Notification System	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Delayed Diagnosis" OR "Diagnoses, Delayed" OR "Diagnosis, Delayed" OR “Errors, Diagnostic” OR “Error, Diagnostic” OR “Missed Diagnosis” OR Misdiagnosis OR Missed Diagnoses OR Misdiagnoses)) AND (((MH Communication OR "Reminder System" OR "Hospital Communication Systems") OR (AB "Reminder System" OR "System, Reminder" OR "Systems, Reminder" OR "Systems, Communication Hospital" OR "Communication Hospital System" OR "Communication Hospital Systems" OR "Hospital System, Communication" OR "Hospital Systems, Communication" OR "System, Communication Hospital" OR "Hospital Communication System" OR "System, Hospital Communication" OR "Communication System, Hospital" OR "Systems, Hospital Communication" OR "Systems, Hospital	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Delayed Diagnosis" OR "Diagnoses, Delayed" OR "Diagnosis, Delayed" OR “Errors, Diagnostic” OR “Error, Diagnostic” OR “Missed Diagnosis” OR Misdiagnosis OR Missed Diagnoses OR Misdiagnoses)) AND ((MH Communication OR "Reminder System" OR "Hospital Communication Systems") OR (AB "Reminder System" OR "System, Reminder" OR "Systems, Reminder" OR "Systems, Communication Hospital" OR "Communication Hospital System" OR "Communication Hospital Systems" OR "Hospital System, Communication" OR "Hospital Systems, Communication" OR "System, Communication Hospital" OR "Hospital Communication System" OR "System, Hospital Communication" OR "Communication System, Hospital" OR "Systems, Hospital Communication" OR "Systems, Hospital

Method	Search	Search String for: CINAHL	Search String for: MEDLINE
• Scientific Integrity Review • Technical Report • Twin Study • Validation Studies CINAHL Publication Types: • Clinical Trial • Corrected Article • Journal Article • Meta-Analysis • Meta Synthesis • Practice Guidelines • Randomized Controlled Trial • Research Review • Systematic Review		Communication" OR "Patient Notification" OR "Automated System" OR "Alert Notification" OR "Critical Test Result" OR "Provider Communication"))	Communication" OR "Patient Notification" OR "Automated System" OR "Alert Notification" OR "Critical Test Result" OR "Provider Communication")))
Search 2008-Present, English Only MedLine Publication Types: • Clinical Trial • Clinical Trial, Phase I • Clinical Trial, Phase II • Clinical Trial, Phase III • Clinical Trial, Phase IV • Comparative Study • Controlled Clinical Trial • Corrected and Republished Article • Evaluation Studies • Guideline • Journal Article • Meta-Analysis • Multicenter Study	Staff Training and Education	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Diagnostic Errors" OR "Error(s), Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis")) AND ((MH "Education, Professional" OR "Simulation Training" OR "Patient Simulation") OR AB (“Education, Professional” OR Training OR “Structured Practice” OR Simulation Training” OR “Patient Simulation”)) AND ((MH Physicians OR "Students, Medical" OR Nursing) OR (AB Physicians OR "Resident Physicians" OR "Medical Students" OR Nursing OR "Healthcare Staff")))	(((MH "Diagnostic Errors" OR "Delayed Diagnosis") OR (AB "Diagnostic Errors" OR "Error(s), Diagnostic" OR Misdiagnosis OR Misdiagnoses OR "Delayed Diagnosis" OR "Missed Diagnosis")) AND (((MH "Education, Professional" OR "Simulation Training" OR "Patient Simulation") OR (AB “Education, Professional” OR Training OR “Structured Practice” OR Simulation Training” OR “Patient Simulation”)) AND ((MH Physicians OR "Students, Medical" OR Nursing) OR (AB Physicians" OR "Resident Physicians" OR "Medical Students" OR Nursing OR "Healthcare Staff")))

Method	Search	Search String for: CINAHL	Search String for: MEDLINE
• Practice Guideline • Published Erratum • Randomized Controlled Trial • Review • Scientific Integrity Review • Technical Report • Twin Study • Validation Studies CINAHL Publication Types: • Clinical Trial • Corrected Article • Journal Article • Meta-Analysis • Meta Synthesis • Practice Guidelines • Randomized Controlled Trial • Research Review • Systematic Review			

AHRQ Pub. No. 20-0029-EF
March 2020